UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIO

TRABAJO FIN DE MÁSTER

CURSO 2016-2017

ANÁLISIS DE SENTIMIENTOS ENFOCADO A LA CALIDAD DE SERVICIO EN LA BANCA ESPAÑOLA

Christian Alejandro Cruz Cornejo

Junio/Septiembre 2017

Tutor: Ramón Alberto Carrasco González

Tabla de contenido

Índice de figuras

1. Introducción

En las condiciones globales actuales, debido a la importancia e influencia de la calidad de los servicios así como a la dificultad de medirlos, algunos investigadores han dedicado mucho tiempo a elaborar instrumentos genéricos para su medición (Barrera, 2012) . En muchos casos, las empresas utilizan metodologías de encuestas diferentes y desarrollan sus propias escalas de medida para medir el mismo problema. Además de las encuestas presenciales tradicionales, las encuestas telefónicas, por correo o fax o las encuestas por Internet. La facilidad con que los cuestionarios en línea pueden ser desarrollados y administrados, junto con la reducción de los costos de las empresas para recopilar datos utilizando estas nuevas herramientas, ha inundado el mercado con encuestas diseñadas para medir la satisfacción del consumidor (Scarilli, 2015) .

Esta amplia variedad de métodos y escalas de medida, hace que sea extremadamente difícil comparar los resultados dentro de la misma empresa o entre empresas competidoras que operan en el mismo sector (Cerezo, 1997), por lo que es necesario utilizar algún tipo de modelo que permita evaluar y medir adecuadamente la calidad de un servicio, además de tener el problema del alto costo que representa realizar encuestas para obtener la información requerida.

Existen modelos clásicos como SERVQUAL (Parasuraman, 1985) que evalúan los factores claves de calidad de servicio prestados el cual considera que todo cliente que adquiere un servicio genera una serie de expectativas que va a recibir a través de distintos canales y una vez recibido hay una serie de factores, dimensiones, que le permite tener una percepción del servicio.

En los últimos años las redes sociales han sido un instrumento importante para obtener información de la opinión de los usuarios hacia un determinado producto o servicio (Queiroz, 2013), cada vez existen más personas que publican sus opiniones o sentimientos habitualmente compuestos de contenido subjetivo y cuyas tendencias pueden ser positivas, negativas o neutrales en función de sus creencias o experiencias por lo que la habilidad de extraer información de datos de redes sociales especialmente de Twitter es un práctica que ya están adoptando organizaciones en todo el mundo sin embargo se enfrentan al inconveniente de analizar un gran volumen de información lo que hace inviable abordar

manualmente el análisis de esas opiniones (Richards, 2015), siendo necesario establecer procesos automáticos que se encarguen de esa tarea en donde el procesamiento de lenguaje natural (PLN) cobra un papel muy importante.

Por ello, en este trabajo se propone medir la calidad de servicio en la Banca Española utilizando la escala SERVQUAL mediante el análisis, clasificación y obtención de sentimientos asociados a características de las opiniones vertidas en la red social Twitter. Esta propuesta se basa en el documento científico *"A model for the integration of e-financial services questionnaires with SERVQUAL scales under fuzzy linguistic modeling"* (Ramón A. Carrasco et. al, 2012).

Este análisis conlleva al uso de técnicas de minería de textos y análisis de sentimientos. Mediante dichas técnicas, las opiniones recogidas dese Twitter dejan de ser datos desestructurados para convertirse en datos estructurados, aptos para su posterior análisis y clasificación.

El presente documento consta de cuatro capítulos adicionales a este. En el siguiente capítulo se hará un repaso del estado del arte de la minería de textos, opiniones y el modelo SERVQUAL. Después en el capítulo tres se fijarán el ámbito y los objetivos del proyecto. Luego, en el capítulo cuatro se describirá cada una de las fases que tiene el proyecto como método de alcance, luego en el capítulo cinco se basará en el desarrollo propiamente del modelo y los algoritmos correspondientes para finalmente, hacer una valoración general del proyecto a modo de conclusiones y futuros trabajos a desarrollar.

2. Estado del Arte

En el presente apartado se presenta los elementos necesarios para entender el propósito del proyecto como es: la Minería de Textos y la escala SERVQUAL.

- **Minería de Textos**

El lenguaje natural que usamos las personas para comunicarnos a menudo es incluido dentro de una categoría denominada "datos no estructurados". En particular, si nos restringimos a la escritura, la minería de textos se refiere al conjunto de procesos que obtiene información

y conocimiento de un conjunto de datos que en principio no tiene orden o no están dispuestos en origen para transmitir esa información. Por lo que se puede decir que la Minería de Texto transforma datos no estructurados a datos estructurados.

Para poder entender este concepto es imprescindible tener claro antes lo que es el la Minería de Datos. Este concepto que surgió hace ya algunos años ayuda a la comprensión y adquisición de conocimiento a partir de la información contenida en bases de datos.

Luis Carlos Molina en su artículo denominado *"Torturando a los datos hasta que confiesen"* define a la minería de Datos como *"la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión"* (Molina, s.f.)

Una vez entendido lo que es la Minería de Datos, podemos extender la misma idea a la Minería de Textos, esta vez los datos a tratar serán, en lugar de los datos de las bases de datos, diversos documentos y textos abarcando tres actividades fundamentales como:

1. Seleccionar los textos pertinentes.
2. Extraer la información o datos clave como características o acontecimientos incluidos en esos textos.
3. Encontrar asociaciones entre esos datos claves

Además existen algunas tareas típicas de la Minería de Textos como (Agirre, 2015):

- Similitud semántica de textos

  Es la medida de la interrelación existente entre dos palabras cualesquiera de un texto, es decir dos palabras por el hecho de tener su existencia en un mismo documento poseen un contexto similar por lo que se entiende que estas están relacionadas.

- Desambiguación del significado

  Aborda el sentido que toma una palabra en una frase cuando la misma puede tener múltiples significados. Por ejemplo la palabra *"banca"* puede tener significados dentro de una oración como "*banca española"* o "*banca para sentarse*".

- Categorización de textos

  Se ocupa de asignar etiquetas que indican a que categoría o categorías corresponde un determinado texto.

  Citando en términos formales; *"existe una función T definida en (DxC) tal que T(di)=ci; donde D es el conjunto de los documentos disponibles, C es el conjunto de*

categorías disponibles, di es un documento cualquiera y ci es el vector de las categorías a las que pertenece el documento di." (Legutier, 2005).

Por tanto el proceso de categorización de textos pretende encontrar una función T' parecido a la función T ya definida.

- Análisis de sentimientos

  Desde el punto de vista de la Minería de Textos, el análisis de sentimientos tiene como objetivo clasificar de forma masiva y automática un grupo de documentos o frases, en función de la polaridad positiva o negativa del lenguaje ocupado en el documento.

  En términos generales, el análisis de sentimiento trata de detectar la actitud del escritor con respecto a posibles emociones, juicios o evaluaciones contenidas en el documento.

  Una tarea básica en el análisis de sentimientos es clasificar la polaridad de un texto, una oración o una característica de manera positiva, negativa o neutra.

La investigación sobre el análisis de sentimientos ha crecido de forma importante desde hace ya algunos años, debido al gran número de campos de aplicación que existen en la actualidad, pero de manera principal por la gran cantidad de opiniones que se pueden recoger en redes sociales, blogs, foros, etc.

Sin embargo, sigue siendo un campo con múltiples retos por resolver. Como por ejemplo la clasificación del sentimiento de un texto como positivo o negativo ya que esto depende, en muchas de las ocasiones, a la forma de cómo se interpreta en función de diversos factores como el cultural o el idioma por mencionar algunos siendo las opiniones o sentimientos subjetivos. Este problema de clasificación, puede tratarse con aprendizaje automático supervisado o no supervisado.

El aprendizaje automático supervisado son aquellos en los que, a partir de un conjunto de ejemplos previamente clasificados, intenta asignar una clasificación a un segundo conjunto de elementos. Su principal ventaja es que se obtiene grandes conjuntos de datos para entrenar un algoritmo debido a que se lo puede realizar de manera masiva. Sin embargo pierde precisión frente a la anotación manual.

Mientras que el aprendizaje automático no supervisado son aquellos en la que no se dispone de un conjunto de datos previamente clasificados, sino que únicamente a partir de las propiedades de los ejemplos intenta agrupar según su *similaridad*[1].

[1] Cualidad de similar. Ejemplo: "Permitirá establecer comparaciones de preferencia o similaridad sobre las categorías." https://es.oxforddictionaries.com/definicion/similaridad

Básicamente, la idea consiste en comparar las palabras consecutivas denominadas *n-gramas* (específicamente bigramas) con distintos patrones sintácticos prefijados.

Un n-grama es una sub secuencia de n elementos de una secuencia dada y que son ampliamente utilizados en tareas de procesamiento de lenguaje natural(Ganesan, 2014). Por ejemplo en el estudio del lenguaje natural se puede construir n-gramas sobre distintos tipos de elementos como son: sílabas, letras, palabras, etc.

Para ciertos valores de n-gramas se tiene sus correspondientes nombres, por ejemplo:

- Los n-gramas de 1 se los denomina unigramas
- Los n-gramas de 2 se los denomina bigramas
- Los n-gramas de 3 se los denomina trigramas

Existen diversas aplicaciones que utilizan n-gramas como:

- Modelos de n-grama

  Se trata de un tipo de modelo probabilístico que permite realizar predicciones estadísticas del próximo elemento de cierta secuencia de elementos. Dichos modelos pueden ser definidos por una cadena de Márkov (UGR, s.f.)[2] de orden n-1

- Técnicas de suavizado

  Las técnicas de suavizado resuelven el problema de la asignación a cero a todos aquellos n-gramas que no aparecen en un conjunto de datos de entrenamiento dados por un determinado modelo estadístico, es decir persigue que todos los n-gramas tengan una probabilidad distinta de cero.

- **Escala SERVQUAL**

La escala SERVQUAL es un instrumento de estudio que pretende medir la calidad del servicio prestado en organizaciones de servicio. Esta escala fue propuesta originalmente por (Parasuraman, 1985) el cual realizó entrevistas a altos ejecutivos de firmas de servicio de atención al cliente y pudo definir la calidad del servicio como una brecha entre las percepciones y las expectativas de los clientes.

[2] Proceso estocástico discreto en el que la probabilidad de que ocurra un evento depende solamente del evento inmediatamente anterior.

Inicialmente la escala proponía diez dimensiones para medir la calidad del servicio. La escala se simplificó más tarde a cinco dimensiones en 1988. En memoria de Parasuraman, Zeithml y Berry el método fue llamado PZB.

En el 2009, Ladhari (R., 2009) hizo una revisión de las diferentes aplicaciones de la escala SERVQUAL entre 1988 a 2008 resaltando la creciente importancia de los servicios en línea. Hasta ese entonces la escala fue creada para medir la calidad del servicio en un contexto de servicio tradicional es decir fuera de línea. Sin embargo con los constantes avances tecnológicos que se desarrollan cada año, la aplicabilidad de esta escala también se ha considerado en entornos de servicios en línea como es el caso del estudio *"A model for the integration of e-financial services questionnaires with SERVQUAL scales under fuzzy linguistic modeling"* (Ramón A. Carrasco et. al, 2012) en el cual se propone utilizar modelos de lógica difusa para medir la calidad del servicio en línea en entidades financieras.

Como se dijo anteriormente, la escala SERVQUAL en la actualidad se compone de cinco dimensiones que miden la calidad del servicio, a continuación se describe cada una de ellas:

- **Fiabilidad**

  Tiene relación con la habilidad de prestar un servicio de forma precisa cuidadosa y fiable. Por ejemplo, que un avión salga y llegue a la hora programada.

- **Tangibilidad**

  Trata sobre la apariencia física de las instalaciones, equipos, empleados y comunicación. Por ejemplo, la limpieza en un restaurante.

- **Capacidad de respuesta**

  Es el deseo genuino de ayudar a los clientes y de servirles de manera rápida. Por ejemplo, La capacidad de una compañía de servicio de Internet para arreglar un problema técnico de manera rápida.

- **Seguridad**

  Se refiere al conocimiento del servicio prestado y amabilidad de los empleados así como su habilidad para transmitir confianza al cliente. Por ejemplo, la confianza que transmite un médico a sus pacientes.

- **Empatía**

  Es el grado de atención personalizada que dispensa la organización a sus clientes. Por ejemplo, el cuidado de la recepción de un hotel por hacer sentir cómodo a su cliente.

- **Cuestionario SERVQUAL**

La puesta en práctica de esta escala consiste en seleccionar una muestra de clientes de la compañía de servicios, los cuales deberán responder a un cuestionario de 22 ítems relacionados con la calidad de servicio recibido.

Estos ítems están distribuidos entre las cinco dimensiones de la siguiente manera:

Ítem 1 a 4 = Tangibilidad

Ítem 5 a 9 = Fiabilidad

Ítem 10 a 13 = Capacidad de Respuesta

Ítem 14 a 17 = Seguridad

Ítem 18 a 22 = Empatía

En la siguiente figura se puede observar el cuestionario simplificado de las 22 preguntas que conforman el cuestionario SERVQUAL.

CALIDAD PERCIBIDA							
1 Los equipos de…….. son aparentemente modernos.	1	2	3	4	5	6	7
2 Las instalaciones de…….. son atractivas visualmente.	1	2	3	4	5	6	7
3 Los empleados de…….. tienen una apariencia correcta.	1	2	3	4	5	6	7
22 Los empleados de…….. comprenden las necesidades específicas de sus clientes.	1	2	3	4	5	6	7

Figura 1. Ejemplo simplificado del cuestionario SERVQUAL de 22 preguntas.

Tomado de: https://rodas5.us.es

3. Ámbito y Objetivos

En el presente apartado se hace una revisión del dominio de actuación de este trabajo, además se detalla sus objetivos principales.

o Ámbito

El ámbito que se ha seleccionado para realizar el presente trabajo sobre Minería de Textos y análisis de sentimientos asociados son las opiniones vertidas en la red Social Twitter sobre la Banca Española. Twitter es una plataforma de comunicación bidireccional con naturaleza de red social que se divide en dos grandes grupos: seguidores ("followers") y seguidos ("followed").

o Objetivos

Objetivo principal

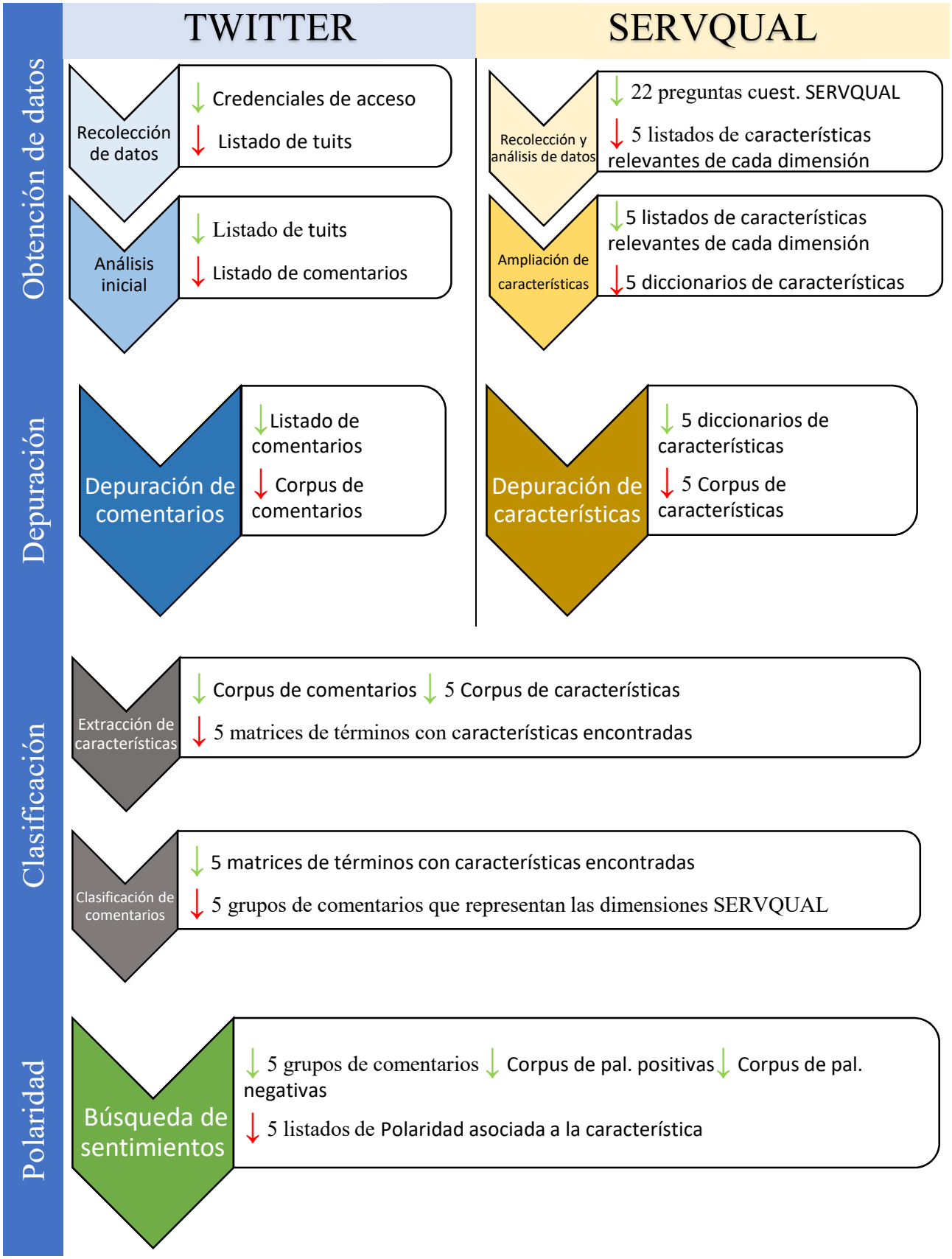
Medir la calidad de servicio que prestan los cinco Bancos Españoles con mayor influencia en las redes sociales mediante un algoritmo que permita utilizar la escala estándar SERVQUAL.

Objetivos específicos

- Clasificar los comentarios obtenidos de Twitter mediante la búsqueda de una característica que permita determinar a cuál de las cinco dimensiones SERVQUAL corresponde.

- Obtener el sentimiento asociado a la característica encontrada en cada comentario mediante técnicas de Minería de Texto.

4. Modelo planteado



↓ ENTRADAS ↓ SALIDAS

A continuación se hará una descripción de cada una de las fases que componen el proyecto

4.1.Fase 1: Obtención de datos

La primera fase del desarrollo del proyecto trata sobre el contacto inicial con los datos a analizar en el cual incluye también su recolección y observación de su calidad. Ésta recolección y análisis de datos se divide en dos grandes grupos de datos a tratar los cuales son:

- Tuits obtenidos de la red social twitter
- Preguntas del cuestionario SERVQUAL.

Twitter

- Recolección de datos

Para la recolección del primer grupo de datos se procederá a obtenerlos mediante uso de credenciales de acceso a las bases de datos de la red social Twitter con la finalidad de establecer una conexión con dicha plataforma e ir realizando diferentes tipos búsqueda de Tuits de acuerdo a parámetros pres establecidos hasta obtener un listado. (Ver proceso en la figura 2).

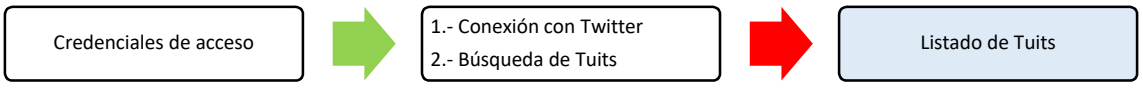


Figura 2. Entradas y salidas de la sub fase recolección de datos de Twitter

- Análisis inicial

Para el análisis inicial del primer grupo, se debe utilizar una muestra aleatoria de Tuits generada a partir del listado obtenido en la sub fase anterior para poder observar de mejor manera las distintas variables y observaciones para obtener los datos más relevantes para el proyecto con la finalidad de obtener un listado de comentarios definitivo a usarse en las siguientes fases. (Ver proceso en la figura 3)

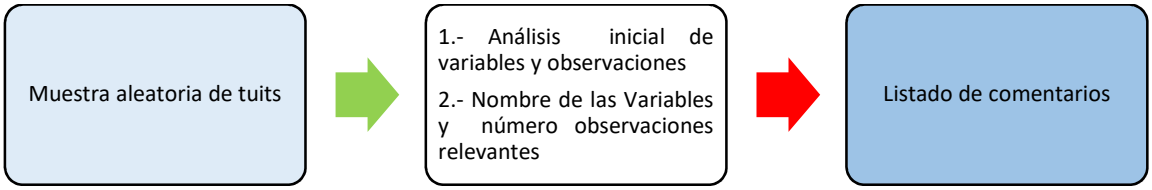


Figura 3. Entradas y salidas de la sub fase análisis inicial de datos de Twitter

SERVQUAL

- Recolección de datos y análisis inicial

Los datos del segundo grupo se obtienen de un listado de 22 preguntas que conforman el cuestionario SERVQUAL divididas en sus respectivas dimensiones para posteriormente analizar cada una de las preguntas del cuestionario y así obtener la o las características más importantes que permitan definir a cada una de las cinco dimensiones. (Ver proceso en la figura 4).

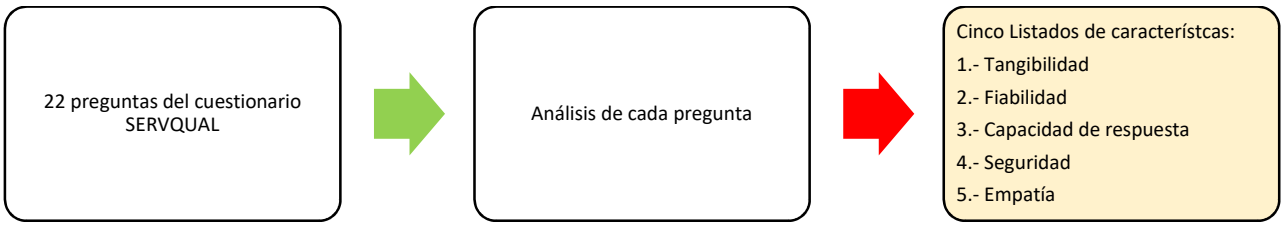


Figura 4 Entradas y salidas de la sub fase recolección y análisis de datos para ServQual

- Diccionario de características

Una vez que se obtiene los cinco listados de características que definen cada dimensión se debe hacer una ampliación de dichos listados para obtener cinco diccionarios de características relevantes más grande que permita describir de mejor manera a cada una de las cinco dimensiones, esto conlleva a realizar una búsqueda de sinónimos de cada característica encontrada en cada listado. (Ver proceso en figura 5).

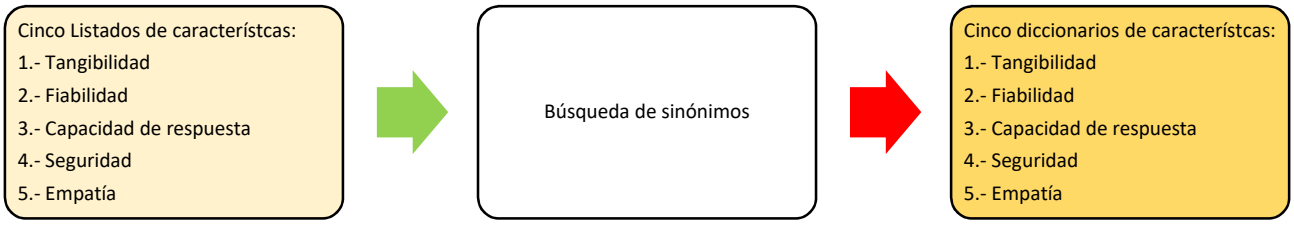


Figura 5. Entradas y salidas de la sub fase ampliación del listado de características

4.2.Fase 2: Depuración de datos

La fase de depuración se centra en utilizar distintas técnicas de Minería de Texto para realizar una limpieza básica de los dos listados de datos obtenidos en la anterior fase. Dichos listados son:

- Listado de comentarios.
- Listado ampliado de características por cada dimensión.

La finalidad de la depuración de datos es obtener los corpus definitivos para ser usados a lo largo del proyecto.

Un corpus[3] es un conjunto de textos, datos u otros materiales sobre determinada materia que pueden servir de base para una investigación o trabajo.

- **Depuración del listado de comentarios**

  La finalidad de esta sub fase es la de crear un corpus de comentarios que será usado a lo largo del proyecto, para esto se utiliza el listado de los comentarios obtenidos y mediante técnicas de Minería de Texto realizar distintas transformaciones y depuraciones como:

  - Remover enlaces a páginas externas (URL).
  - Quitar espacios en blanco.
  - Remover signos de puntuación.
  - Remover números.
  - Transformar a minúsculas.

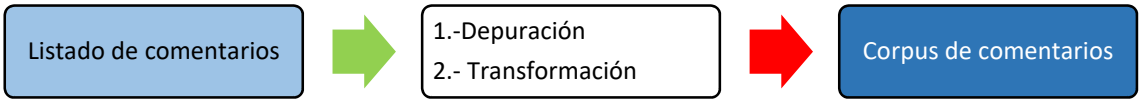


Figura 6 Entradas y salidas de la sub fase depuración de datos de Twitter

- **Depuración de los diccionarios de características**

  Al igual que en el listado de comentarios de Twitter, se debe hacer una depuración y transformación de cada uno de los cinco diccionarios de características utilizando las técnicas de Minería de Texto ya mencionadas y así crear cinco corpus de características que se usarán en todo el proyecto.

  Dichos corpus serán:

  - Corpus para Tangibilidad
  - Corpus para Fiabilidad
  - Corpus para Capacidad de respuesta
  - Corpus para Seguridad
  - Corpus para Empatía

[3] http://www.wordreference.com/definicion/corpus

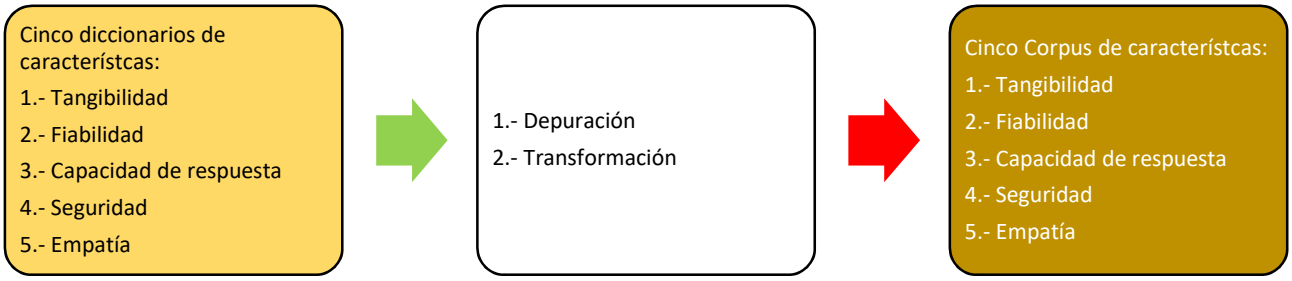


Figura 7. Entradas y salidas de la sub fase depuración de datos ServQual

4.3.Fase 3: Clasificación

La fase de clasificación busca obtener un modelo fiable que sea capaz de distinguir si dentro de un conjunto de comentarios existen algunos que traten sobre la calidad del servicio en la Banca Española y de ser éste el caso que sea capaz de catalogarlo de acuerdo a las cinco dimensiones SERVQUAL.

Para ello se trabaja con los corpus que se crearon en las anteriores fases. Dichos corpus son:

- Corpus de comentarios
- Corpus de características para Tangibilidad
- Corpus de características para Fiabilidad
- Corpus de características para Capacidad de respuesta
- Corpus de características para Seguridad
- Corpus de características para Empatía

Para el corpus de comentarios se implementa nuevas transformaciones de Minería de Textos para maximizar su análisis. Estas nuevas transformaciones son:

- Eliminación de palabras comunes (“Stop Words”)
- Reducción de términos a su raíz (Stemming)

Extracción de características

La finalidad de esta sub fase es obtener un listado de características encontradas en cada comentario que permitan definir si el mismo pertenece o no a alguna de las cinco dimensiones mediante la creación de cinco matrices de términos las cuales cruzan información entre el corpus de comentarios con cada uno de los cinco corpus de dimensión (Ver figuras 8, 9, 10, 11, 12).

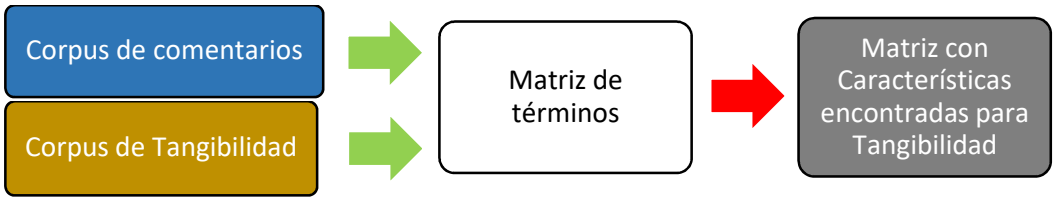


Figura 8 Entradas y salidas de la sub fase extracción de características para Tangibilidad.

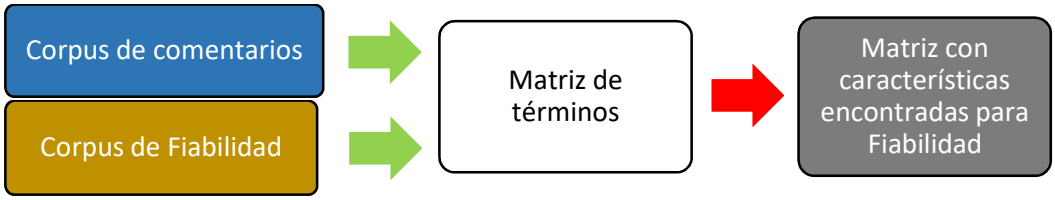


Figura 9 Entradas y salidas de la sub fase extracción de características para Fiabilidad.

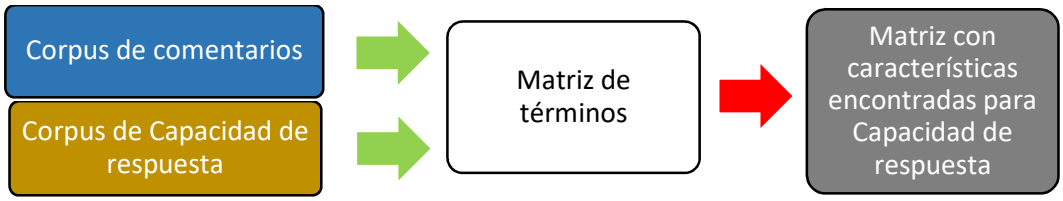


Figura 10 Entradas y salidas de la sub fase extracción de características para Capacidad de Respuesta.

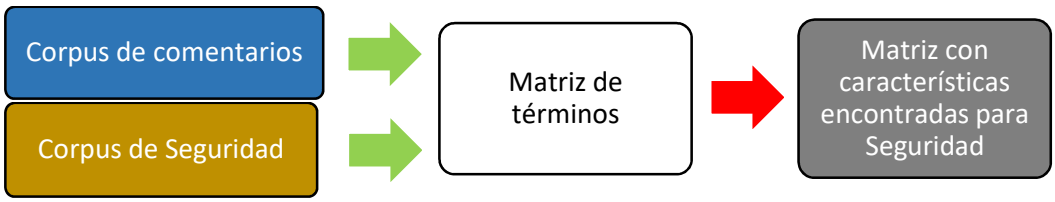


Figura 11 Entradas y salidas de la sub fase extracción de características para Seguridad.

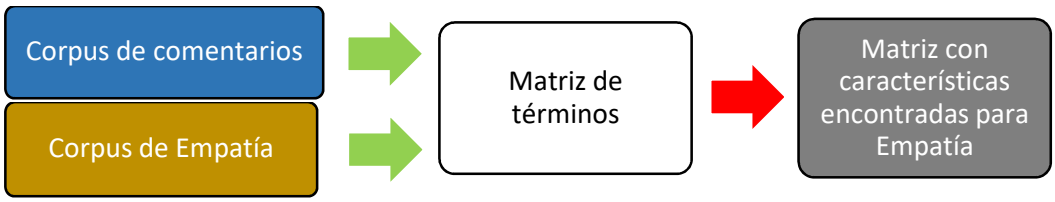


Figura 12. Entradas y salidas de la sub fase extracción de características para Empatía.

Clasificación de comentarios

Para finalizar la fase de clasificación, se utiliza las cinco matrices de términos con las características encontradas en cada comentario y mediante un algoritmo de clasificación se obtiene el grupo de comentarios correspondiente a cada una de las dimensiones SERVQUAL (Ver figuras 13, 14, 15, 16, 17).

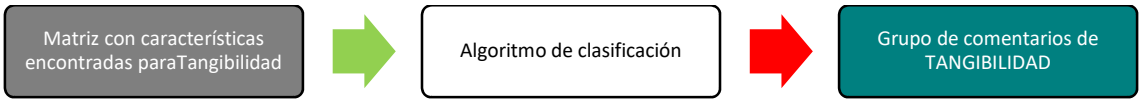


Figura 13 Entradas y salida de la sub fase de clasificación de comentarios para Tangibilidad

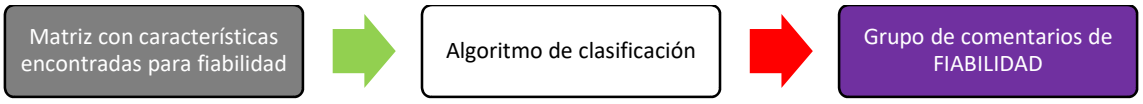


Figura 14. Entradas y salida de la sub fase de clasificación de comentarios para Fiabilidad.

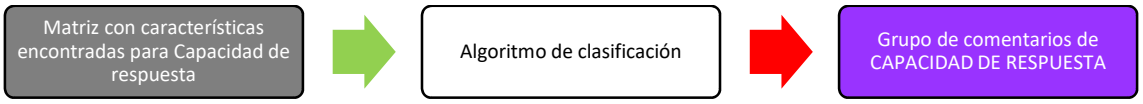


Figura 15. Entradas y salida de la sub fase de clasificación de comentarios para Cap. de Respuesta.

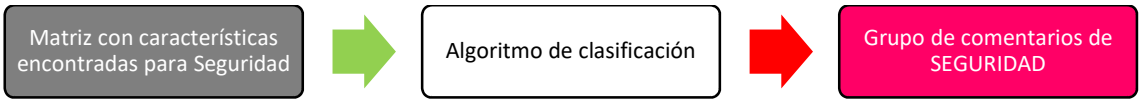


Figura 16. Entradas y salida de la sub fase de clasificación de comentarios para Seguridad.

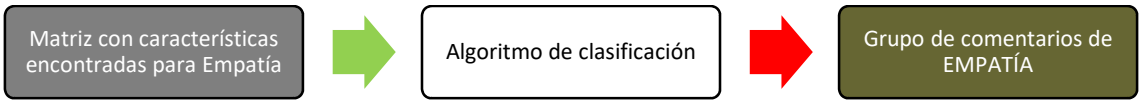


Figura 17. Entradas y salida de la sub fase de clasificación de comentarios para Empatía.

4.4.Fase 4: Polaridad

La finalidad principal de la última fase del proyecto es obtener la polaridad asociada a la característica en cada comentario que determinó si el mismo pertenece a uno de los cinco grupos de dimensiones creados en la fase anterior mediante la búsqueda de sentimientos positivos y negativos.

Para realizar la búsqueda del sentimiento en los comentarios también se utiliza un corpus en español de palabras positivas y otros corpus de palabras negativas previamente creados (Ver figuras 18, 19, 20, 21, 22).

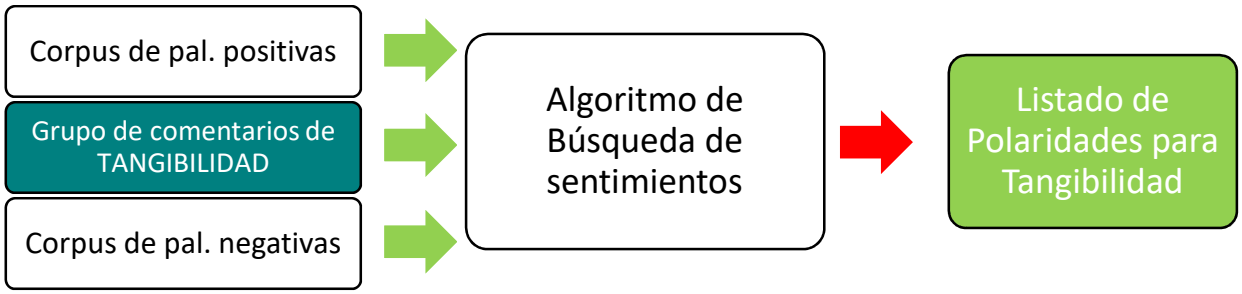


Figura 18. Entradas y salidas de la fase de polaridad para Tangibilidad

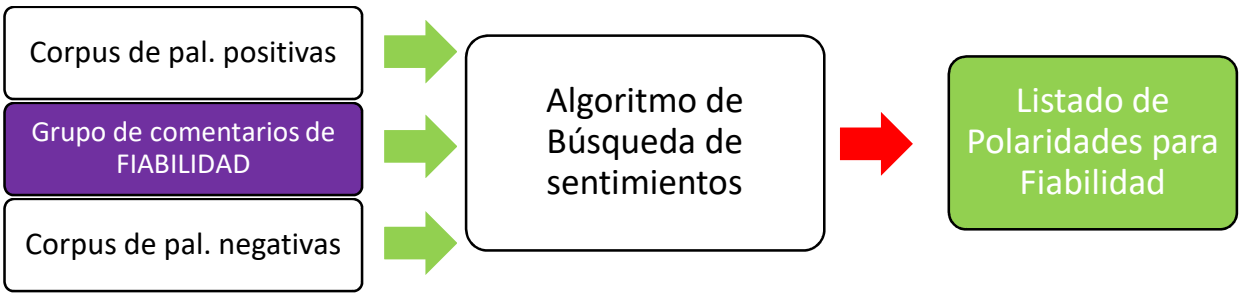


Figura 19. Entradas y salidas de la fase de polaridad para Fiabilidad

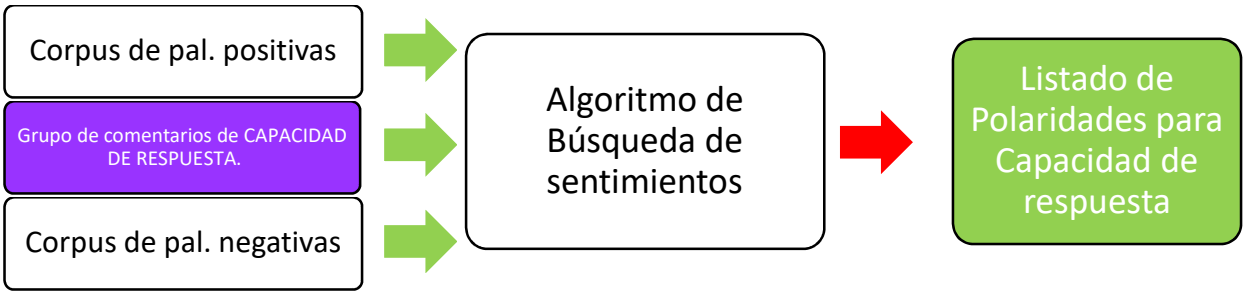


Figura 20. Entradas y salidas de la fase de polaridad para Capacidad de respuesta

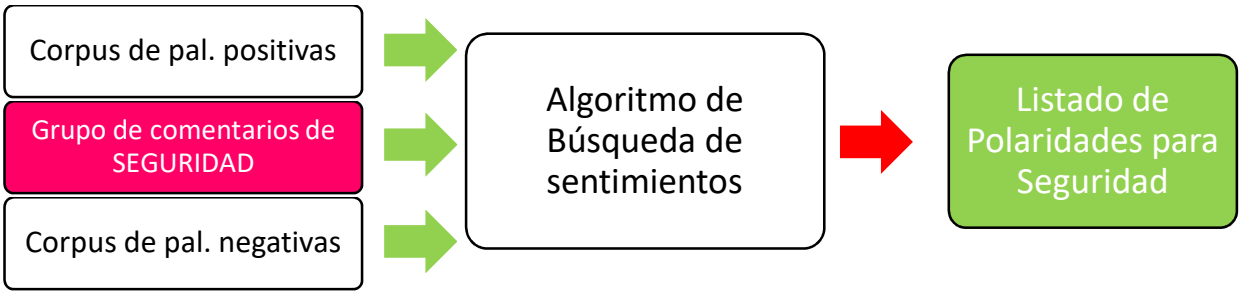


Figura 21. Entradas y salidas de la fase de polaridad para Seguridad.

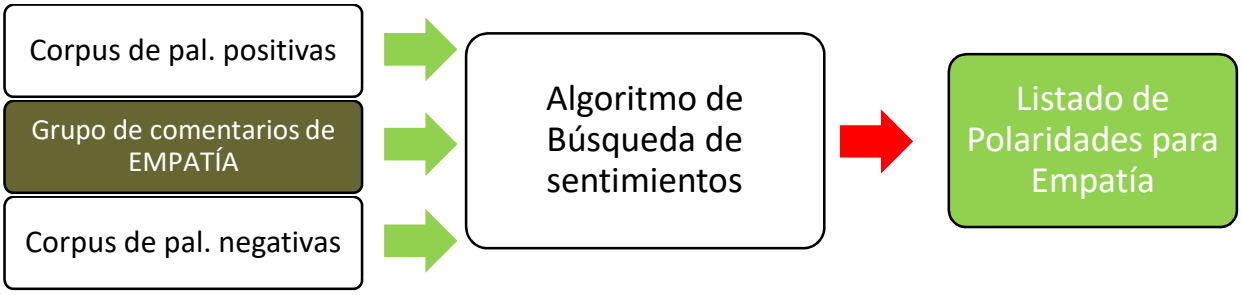


Figura 22. Entradas y salidas de la fase de polaridad para Empatía

5. Desarrollo

A continuación se describen con ejemplos y pruebas las distintas fases del proyecto.

5.1. Obtención de datos

Como se observó en el anterior capítulo se debe hacer dos obtenciones de datos tanto de fuentes como Twitter como fuentes para el cuestionario SERVQUAL.

5.1.1. Twitter

5.1.1.1. Recolección de datos

Nuestra fuente de información principal fue proporcionada por la red social Twitter la cual para acceder a ella se debe seguir una serie de pasos **(ver en anexos)** hasta poder establecer una conexión con sus servidores.

- Conexión con el API de Twitter desde RStudio

Una vez generada las **credenciales de acceso** al API de Twitter, se procede a implementarlas en el aplicativo de RStudio desarrollado utilizando la librería twitterR[4] y usando la función "setup_twitter_oauth" el cual ejecuta la autenticación de Twitter.

Después que se ha establecido la respectiva **conexión con Twitter** se procede a realizar las primeras **búsquedas de Tuits** en lo que tiene que ver con la Banca Española.

[4] https://cran.r-project.org/web/packages/twitteR/twitteR.pdf

- Búsqueda de Tuits en R

Las principales formas de obtención de datos que proporciona Twitter son:

- Rest API: El cual permite realizar búsquedas sobre comentarios, frases, cuentas de usuarios, etc. obteniendo un conjunto de tuits que cumplen las condiciones de la pregunta planteada además de obtener tuits pasados.
- Streaming API: Permite conectar y filtrar los tuits que se están publicando en tiempo real.

En nuestro caso se ha utilizado Rest API ya que permite obtener tuits históricos y además permite realizar búsquedas por múltiples criterios.

Como motivo de prueba las distintas búsquedas se han realizado en lenguaje español y un número de tuits a mostrar de cinco.

Pruebas

Para obtener la mejor parametrización de búsqueda de tuits se han realizado diferentes pruebas con diferentes tipos de parámetros de entrada y se ha evaluado su respectiva salida las cuales se describen a continuación:

Prueba 1:

Criterio a buscar: *"Banco Santander"*

Resultado

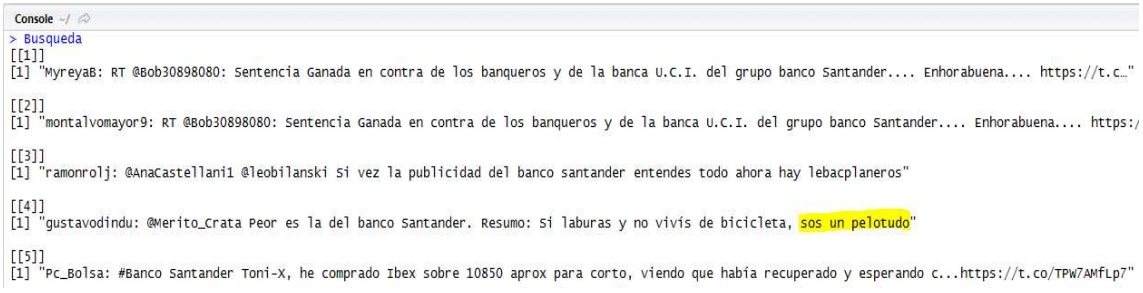


Figura 23. Resultado de la búsqueda por "Banco Santander"

Como se puede observar en la figura 23, al buscar por el nombre del banco *"Banco Santander"* la librería de R ha devuelto tuits que no pueden ser motivos de nuestro estudio ya que pertenecen a zonas fuera de España (Argentina en este caso), por lo que se descarta la primera parametrización y se procede a hacer una búsqueda más específica que trate de delimitar los comentarios vertidos a la Banca en España.

Prueba 2:

Criterio a buscar: *"Banco Santander España"*

Resultado

```
> Busqueda
[[1]]
[1] "BankingFinTech: NEW #JOB: SCF ESPAÑA: Analista de Modelos. Banco Santander [Spain]  #BankingJobs #FinTechJobs https://t.co/yDIJUwoPWS"
```

Figura 24 . Resultado de la búsqueda por "Banco Santander España"

Al implementar a la búsqueda la palabra *"España"*, los resultados arrojados según la figura 24 no fueron los esperados ya que únicamente arrojó un tuit, eliminando posibles comentarios útiles para el análisis, por lo que se ha procedido a realizar una tercera prueba

Prueba 3:

Criterio a buscar: *"#bancosantander"*

Resultado

```
> Busqueda
[[1]]
[1] "Karici: Así la vista desde mi nueva oficina. \n#Godínez #BancoSantander #Workaholic #SantaFe #ViaSantaFe… https://t.co/kx585cwe2u"

[[2]]
[1] "huorellana: #bancosantander no puedo ingresar a la pagina me aparece el siguiente mensaje You don't have permission to access \"https://t.co/MtvZo1SOUq"

[[3]]
[1] "esbolsa: #EncuentroDigital de Ricardo González analizando entre otros #Arkema #BancoSantander #Ibex https://t.co/cEoJkq5FOT https://t.co/TPaZgq6IuO"

[[4]]
[1] "MD_Galdamez: Acto entrega de becas internacionales  2016/2017  #BancoSantander… https://t.co/1KlXvL8iSY"

[[5]]
[1] "JosMariaDez1: RT @bancosantander: #BancoSantander 155 estudiantes de España y América Latina reciben Becas Universidad de Salamanca-Banco Santander https…"
```

Figura 25. Resultado de la búsqueda por el hashtag "#bancosantander"

Al cambiar el criterio de búsqueda por una etiqueta (hashtag[5]) se obtiene un mejor resultado como se observa en la figura 25. Sin embargo se aprecia que cuatro comentarios obtenidos no reflejan ningún criterio de calidad de servicio, por tanto se ha realizado una última prueba

Prueba 4:

Criterio a buscar: *"@santander_es"*

Resultado

```
> Bankinter
[[1]]
[1] "alberto_madalin: @Bota_A_La_Casta @santander_es Argumentalo"

[[2]]
[1] "Bota_A_La_Casta: @alberto_madalin @santander_es Otro cuñao!"

[[3]]
[1] "alberto_madalin: @Bota_A_La_Casta @santander_es Pues haberle enviado dinero por otra vía, los saldos negativos se regularizan cuando… https://t.co/Lyjzo377ZG"

[[4]]
[1] "mama_roja: RT @Bota_A_La_Casta: Cerca de 10.000 personas han visto ya este tuit denunciando las prácticas del Banco Santander @santander_es https://t.…"

[[5]]
[1] "Bota_A_La_Casta: Cerca de 10.000 personas han visto ya este tuit denunciando las prácticas del Banco Santander @santander_es https://t.co/r5ChE6S3MT"
```

Figura 26. Resultado de búsqueda por la cuenta de twitter @santander_es

[5] "palabra o la serie de palabras precedidos por el símbolo de la almohadilla" https://www.significados.com/hashtag/

Como se observa en la figura 26 al realizar una búsqueda por la cuenta oficial del Banco Santander España "@santander_es" los resultados arrojados son más favorables para el análisis de este proyecto aunque cabe decir que existen comentarios "basura" que se deben descartar ya que no aportan ningún valor.

Por tanto, para la búsqueda de comentarios se han usado las cuentas de Twitter oficiales de los 5 bancos de España en estudio.

Dichas cuentas de Twitter se enumeran a continuación:

- @Bankia
- @santander_es
- @BBVA_esp
- @Bankinter
- @BancoSabadell.

Una limitación importante de Rest API es que únicamente permite realizar búsquedas de los últimos 7 días a partir del día en curso, por lo que para solventar este inconveniente se ha hecho un levantamiento de datos de manera diaria en un **periodo comprendido entre el 9 de febrero del 2017 al 13 de marzo del 2017** y almacenándolo en una base de datos NoSQL[6] específicamente Mongo DB[7] obteniendo de esta manera un **listado de 19203 Tuits** agrupados por entidad bancaria (Ver figura 27 y 28).

Col_TodosTweets

_id	empresa	tweets
58a73eb29af7...	@Bankia	[176 elements]
58a73edf9af70...	@Bankia	[151 elements]
58a73f0d9af70...	@Bankia	[179 elements]
58a73f349af70...	@Bankia	[114 elements]
58a73f5d9af70...	@Bankia	[246 elements]
58a73f719af70...	@Bankia	[142 elements]
58a73f999af70...	@Bankia	[291 elements]
58a73fbc9af70...	@Bankia	[572 elements]
58a742149af7...	@santander_es	[29 elements]
58a7422c9af70...	@santander_es	[18 elements]
58a742419af7...	@santander_es	[5 elements]
58a742559af7...	@santander_es	[6 elements]
58a7426e9af7...	@santander_es	[41 elements]
58a7428b9af7...	@santander_es	[133 elements]
58a7429b9af7...	@santander_es	[31 elements]
58a742b29af7...	@santander_es	[20 elements]

Figura 27. Listado de los primeros 15 Tuits almacenados en Mongo DB.

[6] Base de datos no relacional
[7] https://www.mongodb.com/es

	@BancoSabadell	@Bankia	@Bankinter	@BBVA_esp	@santander_es	Total general
Cuenta de Comentario	5268	8999	1606	1539	1791	19203

Figura 28. N° de comentarios obtenidos en el periodo establecido

5.1.1.2. Análisis inicial

Una vez que se ha obtenido un conjunto de comentarios de Twitter se ha procedido a realizar un análisis inicial enfocándose principalmente en las variables y en la calidad de los comentarios.

Para dicho análisis se ha tomado una **muestra aleatoria de 10 tuits** como se observa en la figura.

	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID
1	RT @PattyParadaAgui: Estimad@s amig@s comparto d…	FALSE	0	NA	2017-05-28 20:53:26	FALSE	NA	868933254610640897	NA
2	Estimad@s amig@s comparto denuncia pública relaci…	FALSE	0	NA	2017-05-28 19:58:47	TRUE	NA	868919502184091648	NA
3	RT @CamaraAvila: Hoy lanzamos el Premio #Pymedel…	FALSE	0	NA	2017-05-28 19:12:05	FALSE	NA	868907750637613056	NA
4	@Bota_A_La_Casta @santander_es Argumentalo	FALSE	0	Bota_A_La_Casta	2017-05-28 17:49:53	FALSE	868886784477990912	868887062774206464	324859576
5	@alberto_madalin @santander_es Otro cuñao!	FALSE	0	alberto_madalin	2017-05-28 17:48:47	FALSE	868886395263352834	868886784477990912	2888507410
6	@Bota_A_La_Casta @santander_es Pues haberle envi…	FALSE	0	Bota_A_La_Casta	2017-05-28 17:47:14	TRUE	868879246650900480	868886395263352834	324859576
7	RT @Bota_A_La_Casta: Cerca de 10.000 personas ha…	FALSE	0	NA	2017-05-28 17:28:14	FALSE	NA	868881615916867585	NA
8	Cerca de 10.000 personas han visto ya este tuit den…	FALSE	0	NA	2017-05-28 17:18:50	FALSE	NA	868879246650900480	NA
9	@Superplay360 @santander_es Creo recordar que f…	FALSE	0	Superplay360	2017-05-28 14:13:28	TRUE	868764999925235714	868832599103426562	92330910
10	@santander_es Mentis como bellacos cuando decis …	FALSE	0	santander_es	2017-05-28 14:10:59	FALSE	NA	868831975842476032	1372470686

statusSource	screenName	retweetCount	isRetweet	retweeted	longitude	latitude
<a href="http://twitter.com/download/iphone" rel="…	Loca_Pasion	1	TRUE	FALSE	NA	NA
<a href="http://twitter.com/download/android" rel="…	PattyParadaAgui	1	FALSE	FALSE	NA	NA
<a href="http://twitter.com/download/android" rel="…	valormotivado	3	TRUE	FALSE	NA	NA
<a href="http://twitter.com/download/android" rel="…	alberto_madalin	0	FALSE	FALSE	NA	NA
<a href="http://twitter.com/download/android" rel="…	Bota_A_La_Casta	0	FALSE	FALSE	NA	NA
<a href="http://twitter.com/download/android" rel="…	alberto_madalin	0	FALSE	FALSE	NA	NA
<a href="http://twitter.com/download/android" rel="…	mama_roja	1	TRUE	FALSE	NA	NA
<a href="http://twitter.com/download/android" rel="…	Bota_A_La_Casta	1	FALSE	FALSE	NA	NA
<a href="http://twitter.com/download/android" rel="…	MtqMMEGA	0	FALSE	FALSE	NA	NA
<a href="http://twitter.com/#!/download/ipad" rel="…	linuxantimonopo	0	FALSE	FALSE	NA	NA

Figura 29. Muestra aleatoria del conjunto de tuits.

Al crear la muestra aleatoria de datos se puede observar con mayor claridad todas las variables obtenidas con el API de Twitter y así hacer el **análisis inicial de variables y observaciones** para poder identificar los **nombres de las variables relevantes** al proyecto, dando como resultado la siguiente variable:

- **Text**: Comentario en texto de un determinado usuario

El resto de variables queda fuera del alcance del proyecto, ya que el mismo estará centrado en el análisis de opiniones.

Por tanto, al tener una sala variable en el listado de Tuits se lo ha renombrado como **"Listado de comentarios"**

5.1.2. SERVQUAL

5.1.2.1. Recolección de datos y análisis inicial

Como se explicó anteriormente la fuente de información para SERVQUAL se lo ha obtenido de un **listado de 22 preguntas (Ver anexos)** que describen las cinco dimensiones (tangibilidad, fiabilidad, capacidad de respuesta, seguridad, y empatía) que tiene la escala, por lo que se debió hacer un análisis de cada una de las preguntas para encontrar las características que puedan describir a cada dimensión todo esto basándose en la literatura del presente proyecto en el cual a manera de resumen establece los siguiente:

- Tangibilidad: Trata sobre la apariencia física de las instalaciones, equipos, personal y material de comunicación.

- Fiabilidad: Trata sobre la habilidad para realizar el servicio de modo cuidadoso e íntegro.

- Capacidad de respuesta: Trata sobre la disposición y voluntad para ayudar a los usuarios y proporcionar un servicio rápido

- Seguridad: Conocimientos y atención mostrados por los empleados y sus habilidades para concitar credibilidad y confianza

- Empatía: Atención personalizada que dispensa la organización a sus clientes

Por tanto, haciendo un análisis en las 22 preguntas del cuestionario se ha encontrado las siguientes características que se subrayan a continuación:

- Para tangibilidad

"X cuenta con un equipamiento de aspecto moderno."
"Las instalaciones físicas de X son visualmente atractivas."
"Los empleados de X tienen buena presencia."
"En X, el material asociado con el servicio (como los folletos o los comunicados) es visualmente atractivo."

- Para fiabilidad

"Cuando X promete hacer algo en una fecha determinada, lo cumple"
"Cuando tiene un problema, X muestra un interés sincero por solucionarlo."
"X lleva a cabo el servicio bien a la primera."
"X lleva a cabo sus servicios en el momento que promete que va a hacerlo"
X pone énfasis en unos registros exentos de errores.

- Para capacidad de respuesta

"Los empleados de X le comunican con exactitud cuándo se llevarán a cabo los servicios."
"Los empleados de X le proporcionan un servicio rápido."
"Los empleados de X siempre están dispuestos a ayudarle."
"Los empleados de X nunca están demasiado ocupados para responder a sus preguntas."

- Para seguridad

"El comportamiento de los empleados de X le inspira confianza."
"Se siente seguro en sus transacciones con X."
"Los empleados de X suelen ser corteses con usted."
"Los empleados de X tienen los conocimientos necesarios para contestar a sus preguntas."

- Para empatía

"X le proporciona atención individualizada."
"X tiene unos horarios de apertura o atención adecuados para todos sus clientes."
"X cuenta con unos empleados que le proporcionan una atención personalizada."
"X se interesa por actuar del modo más conveniente para usted."
"Los empleados de X comprenden sus necesidades específicas."

Luego de hacer el **análisis en cada pregunta** y encontrar la característica que lo define se ha obtenido **cinco listados** que son:

Listado 1: Tangibilidad "equipamiento", "instalaciones" , "empleados", "material" ,"servicio", "folleto", "comunicados"
Listado 2: Fiabilidad "promete", "fecha", "problema", "solucionarlo", "servicio", "primera", "servicios", "énfasis", "registros", "errores"
Listado 3: Capacidad de respuesta "empleados", "exactitud", "servicios", "empleados", "servicio", "empleados", "dispuestos ", "empleados", "ocupados", "preguntas."
Listado 4: Seguridad "comportamiento", "empleados", "seguro", "transacciones", "empleados", "empleados", "conocimientos", "preguntas"
Listado 5: Empatía "atención", "individualizada", "horarios", "apertura", "atención", "clientes", "empleados", "atención", "personalizada", "interesa", "conveniente", "empleados", "necesidades", "específicas"

Tabla. Cinco listados de características encontradas en el cuestionario SERVQUAL

5.1.2.1.1. Diccionario de características

Una vez que se ha realizado el análisis y la obtención de **cinco listados con las características relevantes** de cada dimensión SERVQUAL se lo ha ampliado mediante la **búsqueda de sinónimos (ver anexos)** obteniendo así **cinco diccionarios de palabras** que se enumeran a continuación:

Diccionario 1: Tangibilidad

- *"equipamiento"*
  Sinónimos: Suministro, aprovisionamiento, infraestructura
- *"instalaciones"*
  Sinónimos: Construcción, emplazamiento, establecimiento
- *"empleados"*
  Sinónimos: Dependiente, funcionario, auxiliar, subalterno, trabajador, asalariado
- *"material"*
  Sinónimos: Elemento, componente
- *"servicio"*
  Sinónimos: Prestación, trabajo, asistencia, función, oficio ocupación, ayuda
- "folleto"
  Sinónimos: Impreso, fascículo, cuaderno, encarte, revista, panfleto, catálogo
- *"comunicados"*
  Sinónimos: Aviso, mensaje, oficio, notificación, comunicación, despacho

Diccionario 2: Fiabilidad

- *"promete"*
  Sinónimos: Proponer, ofrecer, convenir, garantizar, comprometerse
- *"fecha"*
  Sinónimos: Momento, plazo, tiempo
- *"problema"*
  Sinónimos: Duda, pregunta, cuestión, incógnita, dilema, ejercicio, enigma
- *"solucionarlo"*
  Sinónimos: Solventar, resolver, reparar, arreglar, componer, remediar
- *"servicio"*(x2)[8]
  Sinónimos: Prestación, trabajo, asistencia, función, oficio ocupación, ayuda
- *"primera"*
  Sinónimos: Primeramente, antes, previamente
- *"énfasis"*
- Sinónimos: Intensidad, vigor, energía, vehemencia, fuerza, realce
- *"errores"*
  Sinónimos: Confusión, equivocación, errata, fallo, falta, disparate

Diccionario 3: Capacidad de respuesta

- *"empleados"*(x4)
  Sinónimos: Dependiente, funcionario, auxiliar, subalterno, trabajador, asalariado

[8] X2: número de veces que se repite la característica en una dimensión ServQual.

• *"exactitud"* Sinónimos: Justeza, precisión, puntualidad, fidelidad • *"servicios"*(x2) Sinónimos: Prestación, trabajo, asistencia, función, oficio ocupación, ayuda • *"preguntas"* Sinónimos: Interrogación, interrogante, interpelación, cuestión, consulta
Diccionario 4: Seguridad • *"comportamiento"* Sinónimos: conducta • *"empleados"* (x3) Sinónimos: Dependiente, funcionario, auxiliar, subalterno, trabajador, asalariado • *"transacciones"* Sinónimos: Trato, intercambio, acuerdo, convenio, negocio, negociación • *"conocimientos"* Sinónimos: Inteligencia, discernimiento, consciencia, razón, intuición • *"preguntas"* Sinónimos: Interrogación, interrogante, interpelación, cuestión, consulta
Diccionario 5: Empatía • *"atención"*(x3) Sinónimos: Beneficencia, filantropía, caridad, ayuda, auxilio "individualizada" Sinónimos: Particular, propio, personal, peculiar, privado, especial • *"horarios"* Sinónimos: Itinerario, programa, agenda • *"apertura"* Sinónimos: Iniciación, comienzo, principio • *"clientes"* Sinónimos: Parroquiano, asiduo, comprador, consumidor, usuario • *"empleados"*(x3) Sinónimos: Dependiente, funcionario, auxiliar, subalterno, trabajador, asalariado • *"personalizada"* Sinónimos: Individualizada, particular, propi, personal, peculiar, privado • *"específicas"* Sinónimos: Determinado, especial, particular, peculiar, propio, concreto

Tabla. Diccionarios de características por cada dimensión ServQual.

5.2. Depuración de los datos

5.2.1. Depuración del listado de comentarios

En la anterior fase se obtuvo un **listado de 19203 comentarios** el cual tiene una serie de datos que no aportan valor al proyecto. Estos datos son los denominados "basura" por lo que debe ser depurado mediante distintas técnicas de Minería de Texto.

Al tener un número significativo de observaciones se ha tomado una muestra aleatoria del listado de comentarios para poder analizar y aplicar una serie de operaciones básicas de limpieza a los textos para facilitar su uso en las siguientes fases. Dicha muestra se la observa en la figura 30.

```
 [1] "RT @martu_ky: @AmaliaBlanco2 @Bankia Os podrá gustar más o menos Bankia pero lleva unos años esforzándose mucho en hacer las cosas bien, co…"
 [2] "RT @APADISASOC: Tienes ya tu dorsal? Corre a porel,nos los quitan de las manos;) @Cofares @RisiOficial @Unicampus @Bankia @Sansensayalocal…"
 [3] "RT @PAH_Madrid: #MonicaSeQueda @Bankia y el juzg. 31 intentan por 2ª vez desahuciar a esta familia, pese a que tiene derecho a la moratoria…"
 [4] "RT @APADISASOC: Tienes ya tu dorsal? Corre a porel,nos los quitan de las manos;) @Cofares @RisiOficial @Unicampus @Bankia @Sansensayalocal…"
 [5] "RT @APADISASOC: Tienes ya tu dorsal? Corre a porel,nos los quitan de las manos;) @Cofares @RisiOficial @Unicampus @Bankia @Sansensayalocal…"
 [6] "@Bankia @fedetoempresas La combinación de las 2, y la omnicanalidad nos harán más fuertes."
 [7] "RT @Bankia: #Goirigolzarri ante @fedetoempresas: \"Por mucha #digitalización a la que asistamos, el cliente demandará una relación personal…"
 [8] "RT @APADISASOC: Tienes ya tu dorsal? Corre a porel,nos los quitan de las manos;) @Cofares @RisiOficial @Unicampus @Bankia @Sansensayalocal…"
 [9] "RT @Bankia: Conoce el #CampusBankiadelEsfuerzo: valores y sueños por medio del baloncesto, gracias a @valenciabasket #BlogBankia https://t.…"
[10] "RT @PAH_Madrid: #MonicaSeQueda @Bankia y el juzg. 31 intentan por 2ª vez desahuciar a esta familia, pese a que tiene derecho a la moratoria…"
```

Figura 30. Listado de los 10 primeros comentarios de la muestra aleatoria del listado.

Con los diez comentarios aleatorios obtenidos del listado de inicio se ha podido identificar las operaciones de limpieza que se deben implementar (Ver figura 31).

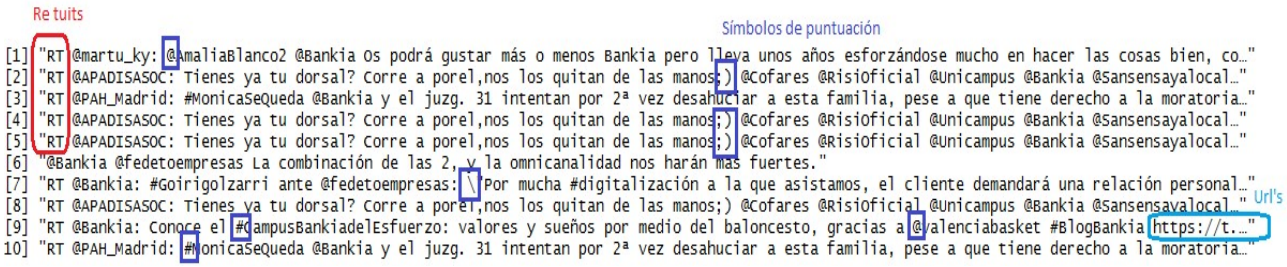


Figura 31. Identificación de operaciones de limpieza básica.

Por tanto, se ha procedido a limpiar todo el listado de comentarios usando la librería tm[9] de R que permite **depurar y transformar** texto para mejorar su contenido, por lo que se han realizado operaciones de:

- Quitar Re tuits (RT)
- Remover enlaces a páginas externas (Url)
- Quitar espacios en blanco
- Remover signos de puntuación
- Transformar a minúsculas

[9] https://cran.r-project.org/web/packages/tm/tm.pdf

```
 [1] "rt martuky amaliablanco bankia os podrá gustar más o menos bankia pero lleva unos años esforzándose mucho en hacer las cosas bien co…"
 [2] "rt apadisasoc tienes ya tu dorsal corre a porelnos los quitan de las manos cofares risioficial unicampus bankia sansensayalocal…"
 [3] "rt pahmadrid monicasequeda bankia y el juzg  intentan por  vez desahuciar a esta familia pese a que tiene derecho a la moratoria…"
 [4] "rt apadisasoc tienes ya tu dorsal corre a porelnos los quitan de las manos cofares risioficial unicampus bankia sansensayalocal…"
 [5] "rt apadisasoc tienes ya tu dorsal corre a porelnos los quitan de las manos cofares risioficial unicampus bankia sansensayalocal…"
 [6] "bankia fedetoempresas la combinación de las  y la omnicanalidad nos harán más fuertes"
 [7] "rt bankia goirigolzarri ante fedetoempresas por mucha digitalización a la que asistamos el cliente demandará una relación personal…"
 [8] "rt apadisasoc tienes ya tu dorsal corre a porelnos los quitan de las manos cofares risioficial unicampus bankia sansensayalocal…"
 [9] "rt bankia conoce el campusbankiadelesfuerzo valores y sueños por medio del baloncesto gracias a valenciabasket blogbankia httpst…"
[10] "rt pahmadrid monicasequeda bankia y el juzg  intentan por  vez desahuciar a esta familia pese a que tiene derecho a la moratoria…"
```

Figura 32. Limpieza inicial del conjunto de datos

Una vez depurado el listado se ha procedido a crear el **corpus de comentarios** mediante código de R.

5.2.2. Depuración del diccionario de características

Al igual que sucedió con el listado de comentarios, se ha realizado una **depuración y transformación** en cada uno de los **cinco diccionarios de características** obteniendo de esta manera **cinco corpus** que describen cada dimensión SERVQUAL. Las operaciones realizadas fueron:

- Quitar espacios en blanco
- Remover signos de puntuación
- Transformar a minúsculas

5.3.Clasificación

5.3.1. Extracción de características

Como se dijo anteriormente, la finalidad de este apartado es obtener un listado de características encontradas en cada comentario el cual nos permite saber si pertenece o no a alguna dimensión, para ello debemos cruzar información entre el corpus de comentarios con cada uno de los corpus de dimensiones. Esto se lo ha realizado mediante la implementación de una matriz de términos[10].

Se trata de una matriz de dimensión mxn, donde m es el número de comentarios a procesar y n es el número de características existentes en dichos comentarios. Los valores de la matriz son el número de veces de cada **comentario (fila)** contiene la **característica (columna)** dada.

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	…
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	

Figura 33. Ejemplo de una matriz de términos. Obtenida de https://nlp.stanford.edu

[10] https://mukom.mondragon.edu/ict/mineria-de-textos-para-la-clasificacion-de-documentos-en-espanol-con-r/

5.3.1.1. Matriz de términos

Como se observa en la figura 33 la matriz de términos permite encontrar asociaciones entre filas y columnas. En el caso del actual proyecto las filas representan los comentarios y las columnas las características de cada dimensión.

Ejemplo:

	infraestructura	instalaciones	oficinas	establecimiento	...
"Las oficinas son muy confortables en Bolaños de Calatrava"	0	0	1	0	0

Tabla. Ejemplo de matriz de términos para tangibilidad

Para crear las cinco diferentes matrices de términos se ha usado "TermDocumentMatrix"[11] que pertenece a la librería "tm" de R.

Como se explicó anteriormente, se debe realizar un cruce de información entre los distintos corpus obteniendo las siguientes variables como resultados:

- Corpus de comentarios con Corpus de características para tangibilidad

vist	ayud	trabaj	emple	funcion	pint	equip	ocup	car
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Tabla. Resultado de los primeros registros del cruce de información para Tangibilidad

- Corpus de comentarios con Corpus de características para fiabilidad

error	oblig	resolv	ayud	funcion	solucion	asegur	dud	plaz	falt
1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Tabla. Resultado de los primeros registros del cruce de información para Fiabilidad

[11] https://www.rdocumentation.org/packages/tm/versions/0.6-2/topics/TermDocumentMatrix

- Corpus de comentarios con Corpus de características para Capacidad de respuesta

ayud	trabaj	aleg	emple	apoy	funcion	dud	consult	continu	respond
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Tabla. Resultado de los primeros registros del cruce de información para Capacidad de respuesta

- Corpus de comentarios con Corpus de características para seguridad

compromis	trabaj	emple	dud	consult	segur	negoci	respond	inspir	pregunt
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Tabla. Resultado de los primeros registros del cruce de información para Seguridad

- Corpus de comentarios con Corpus de características para empatía

client	entreg	especial	ejerc	jorn	agend	ejecut	horari	capt	program
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Tabla. Resultado de los primeros registros del cruce de información para Empatía

Una vez que se creado las distintas matrices de términos se ha procedido a implementar en cada una de ellas una nueva variable denominada "Características_encontradas" en la cual se almacena la característica o las características encontradas en el comentario. Cabe decir que se ha creado una función que permite extraer el conjunto de características de los comentarios.

Ejemplo:

	Carac. Encontrada	infraestructura	instalaciones	oficinas	…
"Las oficinas son muy confortables en Bolaños de Calatrava"	oficinas	0	0	1	0

Tabla. Ejemplo de Matriz de términos ampliado con variables para clasificación de dimensión Tangibilidad.

5.3.2. Pruebas

A continuación se presenta un conjunto de pruebas para dicho mejoramiento.

Las pruebas correspondientes a la extracción de características se lo han realizado tomando como muestra únicamente el **corpus de Tangibilidad,** posterior a tener el modelo adecuado, se lo replicará a los cuatro corpus restantes.

Como método de evaluación de cada prueba se mide el número de palabras frecuentes y características encontradas en la totalidad del corpus mediante gráfica de barras.

El número de palabras frecuentes sirve para mejorar el modelo en la fase de clasificación de comentarios, por lo que se busca que no exista dispersión de términos.

Prueba 1
Entradas: • Corpus de comentarios • Corpus de dimensión Servqual TANGIBILIDAD

Resultados

Términos Frecuentes

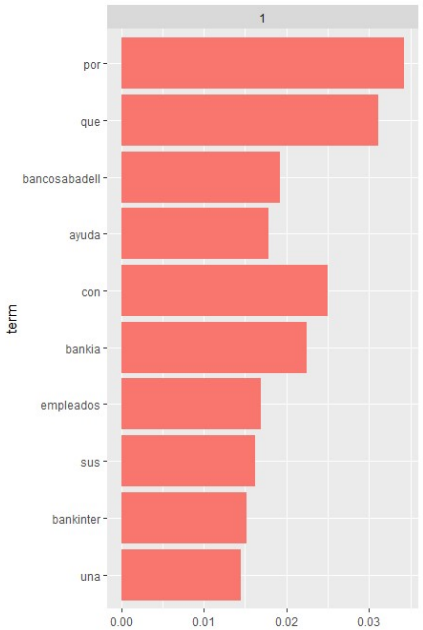


Figura 34. Palabras frecuentes de la prueba 1

Características encontradas

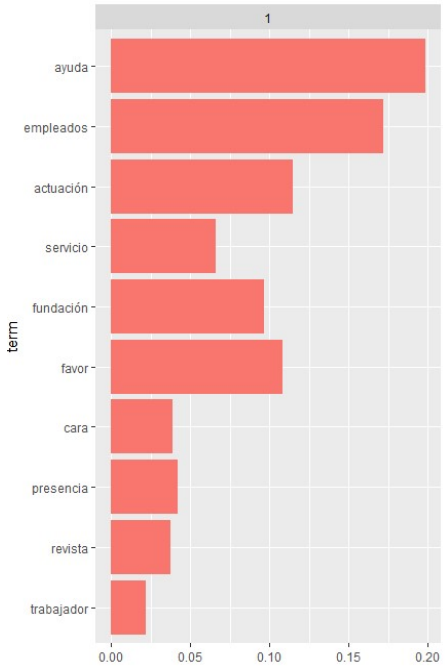


Figura 35. Características encontradas con la prueba 1

Evaluación

Como se puede observar en la figura 35, el modelo ha identificado algunas características dentro del corpus de comentarios y los ha almacenado en la matriz de términos, sin embargo se aprecia en la figura 34 que existen muchas palabras que no aportan valor al modelo como por ejemplo "por", "que". Estas palabras se las denomina Stop Word.

Stop Words o palabras vacías[12] se refiere a todas aquellas palabras que carecen de un significado por si solas. Comúnmente suelen ser artículos, preposiciones, conjunciones, pronombres, etc.

Ejemplo:

"Bankia el señor que teneis por las tardes en Luis se Hoyos Sainz es un borde penoso el servicio que ofrece Humillante e inhumano"

Stop Words encontradas:
"el" (x2), *"que"*(x2), *"por", "las", "en", "se", "es", "un", "e"*

Eliminación de Stop Words
"Bankia señor teneis tardes Luis Hoyos Sainz borde penoso servicio ofrece Humillante inhumano".

En la siguiente figura de nube de términos se los puede apreciar de mejor manera los Stop Words encontrados en el corpus de tangibilidad.

Figura 36. Nube de palabras de la prueba uno

[12] http://www.vozidea.com/que-son-las-stop-words-o-palabras-vacias

Al tener una gran cantidad de palabras vacías se hace imprescindible eliminarlas para mejorar el modelo. Por tanto como objeto de análisis y explicación se ha obtenido del corpus de comentarios de una muestra aleatoria de 400 observaciones para eliminar las palabras vacías y a posteriori implementarlo en todo el corpus.

Obtención de palabras frecuentes

El primer paso para depurar la muestra aleatoria es obtener las palabras más frecuentes encontradas en dicha muestra como se muestra en la figura 37.

	word	freq
bankia	bankia	426
que	que	133
amaliablanco	amaliablanco	114
los	los	96
por	por	68
con	con	60
una	una	49
las	las	43
fantÃ	fantÃ	36
para	para	35
soluciÃ	soluciÃ	33
stica	stica	31
somosabanca	somosabanca	30
muchas	muchas	29
espaÃ	espaÃ	29
gracias	gracias	28
sfe	sfe	28
tiene	tiene	26
mikihoyos	mikihoyos	25
facua	facua	24

Figura 37. Palabras frecuentes muestra aleatoria de 400 comentarios

Eliminación de palabras comunes o vacías (Stop Words)

El segundo paso es el análisis de palabras vacías a eliminar. Estas se las puede observar en la siguiente nube de palabras:

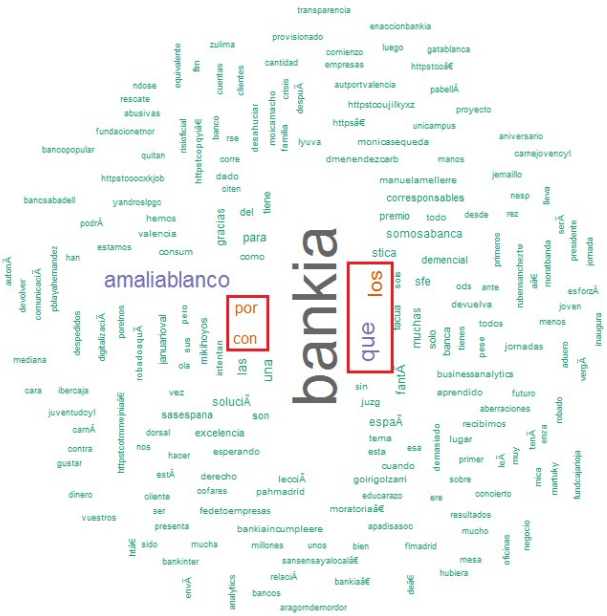


Figura 38. Nube de palabras de la muestra aleatoria de comentarios

Por tanto como último paso, se ha procedido a eliminar las palabras vacías en todo el corpus de comentarios implementando un listado de dichas palabras en español **(ver anexos).**

	word	freq
bankia	bankia	426
amaliablanco	amaliablanco	114
fantÃ	fantÃ	36
soluciÃ	soluciÃ	33
stica	stica	31
somosabanca	somosabanca	30
espaÃ	espaÃ	29
gracias	gracias	28
sfe	sfe	28
mikihoyos	mikihoyos	25
facua	facua	24
premio	premio	22
sasespana	sasespana	22
banca	banca	21
excelencia	excelencia	21
corresponsables	corresponsables	21
demencial	demencial	20
januarioval	januarioval	20
tema	tema	19
esperando	esperando	18

Figura

Figura 39 . Resultado de palabras frecuentes de la muestra aleatoria sin stop words.

Prueba 2
Entradas: • Corpus de comentarios sin Stop Words • Corpus de dimensión Servqual TANGIBILIDAD

Resultados

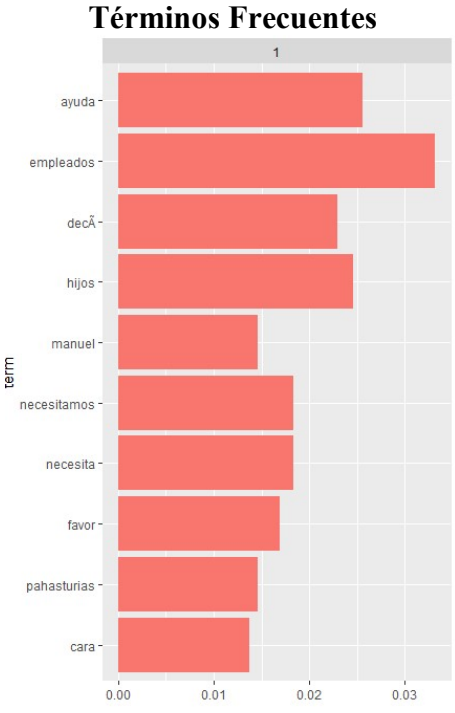


Figura 40. Palabras frecuentes con la prueba 2

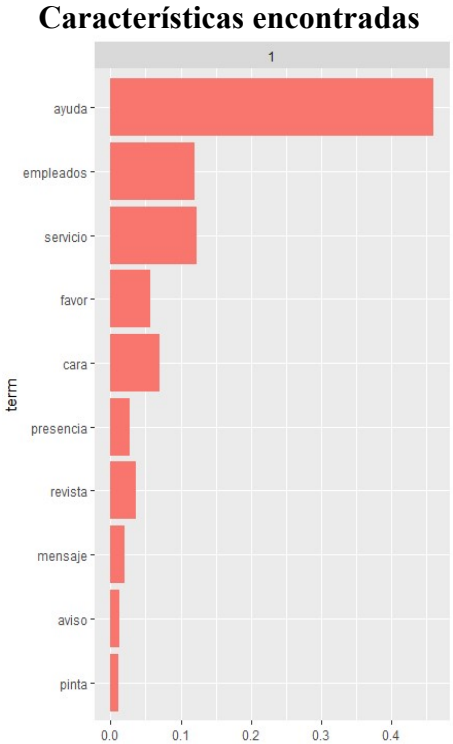


Figura 41. Características encontradas con la prueba 2

Caracteristicas_Encontradas
aviso
facha
favor
favor,ayuda
servicio

Tabla. Características almacenadas en la matriz de términos para Tangibilidad con la prueba 2

Evaluación

Como se puede observar en la tabla de la prueba 2, el modelo ha mejorado al quitar las palabras vacías ya que ha identificado más de una característica en un comentario *("Favor y ayuda").*

Sin embargo se ha identificado un problema al ver los resultados de la figura 40 de términos frecuentes ya que existen palabras con un mismo término de raíz como por ejemplo *"necesitamos"* y *"necesita"*

Reducción de términos a su raíz

La reducción de términos a su raíz[13] (Stemming) implica cortar los afijos y dejar únicamente sus raíces.

Ejemplo:

Término: *"bancario"*

Raíz: banca

Afijo: rio

Por tanto "bancario" y "banca" podrán ser tratados como un solo término.

Pruebas

Nuevamente obtenemos una muestra aleatoria de 400 comentarios en los cuales se debe implementar la reducción de términos como se observa en la figura 42.

[13] http://www.cemla.org/PDF/boletin/PUB_BOL_LXII-03-01.pdf

abancapero	1
abengoa	1
aberraciones	11
abona	2
abrir	2
abusado	3
abusiva	2
abusivas	7
abuso	2
acaban	1
acabo	1

Figura 42. Términos susceptibles de reducción a su raíz

La figura 42 muestra la existencia de diferentes palabras con el mismo sentido en la oración como son los casos de *"abusado"*(x3), *"abusiva"* (x2), *"abusivas"* (x7), *"abuso"* (x2).

Mediante el uso de la librería SnowballC[14] se ha reducido a su raíz los términos de cada uno de los comentarios de la muestra del corpus de comentarios como se observa en la figura 43.

word	freq
abancap	1
abengo	1
aberr	11
abon	2
abrir	2
abus	14
acab	2

Figura 43. Reducción de términos a su raíz.

Cabe decir que este procedimiento se lo ha replicado en todo el corpus de comentarios.

Prueba 3
Entradas: • Corpus de comentarios sin Stop Words y con Steaming • Corpus de dimensión Servqual TANGIBILIDAD

[14] https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf

Resultados

Términos Frecuentes

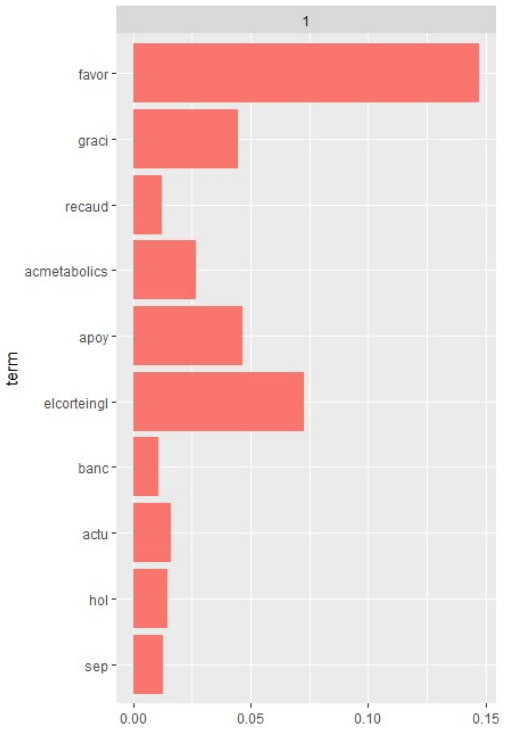


Figura 44. Palabras frecuentes con la prueba 3

Características encontradas

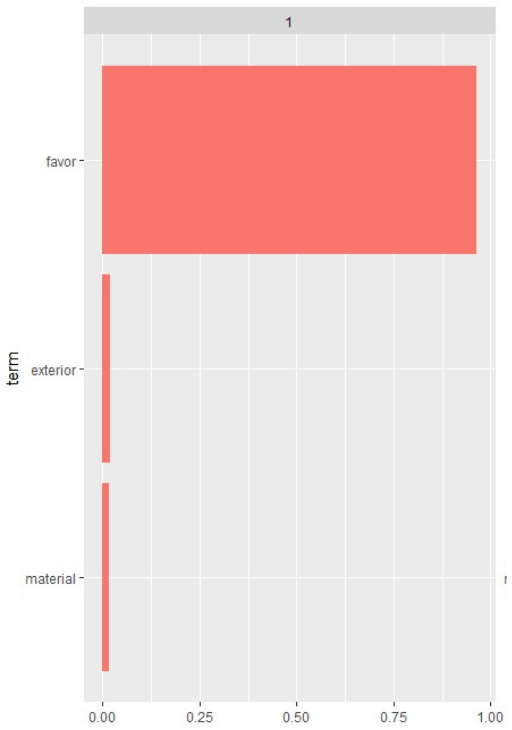


Figura 45. Características encontradas con la prueba 3

Evaluación

En la tercera prueba del modelo se observa que las palabras frecuentes con su misma raíz se han unificado y de esta manera se ha reducido su número.

Sin embargo en cuanto a las características encontradas el resultado no ha sido el esperado ya que ha identificado pocas de ellas. Esto se debe a que al estimizar [15] los comentarios, el algoritmo de búsqueda pierde fuerza al no tener la palabra completa. Por lo tanto se ha debido estimizar también los corpus de las dimensiones SERVQUAL.

Prueba 4
Entradas: • Corpus de comentarios sin Stop Words y con Stemming • Corpus de dimensión SERVQUAL TANGIBILIDAD con stemming

[15] Conversión de una palabra a su raíz.

Resultados

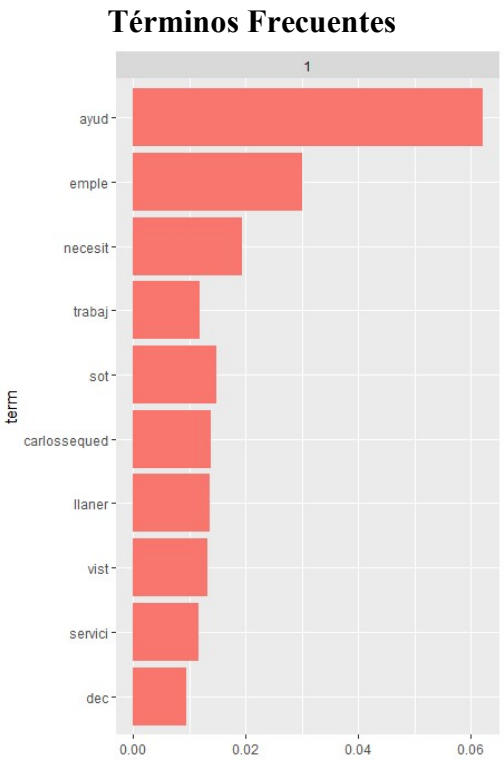


Figura 46. Palabras frecuentes con la prueba 4

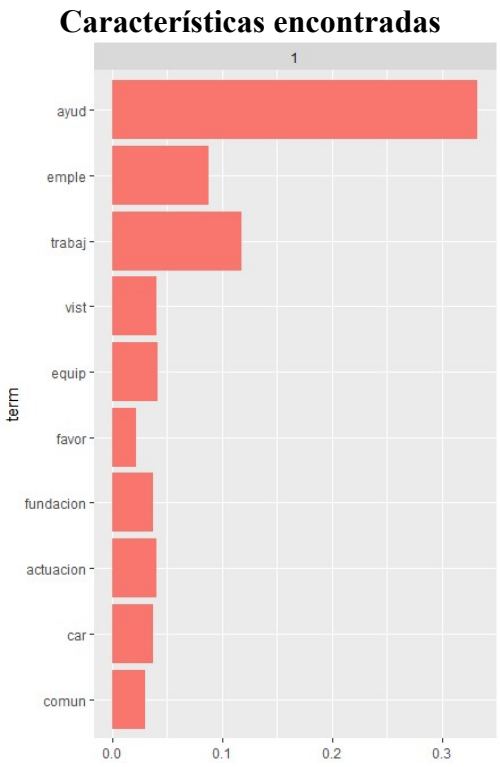


Figura 47. Características encontradas con la prueba 4

Caracteristicas_Encontradas
avis
avis,favor
vist
mensaj
favor

Tabla. Cinco primeros comentarios obtenidos con la prueba 4

Evaluación

Los resultados de la prueba 4 han mejorado de manera considerable ya que el modelo ha sido capaz de encontrar más de una característica asociada a un comentario, además de minimizar el número de palabras frecuentes dentro de la dimensión en análisis.

Para finalizar, se ha replicado el modelo obtenido en la prueba 4 en los cuatro corpus restantes y así seguir evaluando el modelo con datos desconocidos.

- Resultados en la Matriz de características para TANGIBILIDAD

Comentario	**Car. encontrada**
" no fiarse de bankia ha engañado a sus clientes estafado a sus propios accionistas y mentido a sus empleado"	emple
"casidios rennoaltocampoo bankia nasí funcionan los sindicalistas vergonzoso país"	funcion
" si quieres saber qué pintamos en snapchat síguenos bankiatecuenta"	pint
"por las tarjetas black y por el ere ejercido sobre las trabajadoras de bankia se merecen una dura condena txellgl"	trabaj
" en la intranet nos han puesto calculadora para q cada empleado de bankia podamos calcular ntro impacto ambiental en lo"	emple

Tabla. Matriz de características con cinco comentarios aleatorios para TANGIBILIDAD

- Resultados en la Matriz de características para FIABILIDAD

Comentario	**Car. encontradas**
"bankia como siempre no cambian error estrellados"	error
"lim obligado a conservar mestalla en buen estado si no lo hace bankia resolvería el préstamo de millones"	oblig,resolv
"buscando una solución para erikasequedasindeuda en bankia de soto"	solucion
"quereis engañarme quereis joderme con preferentes quereis mucho a rodrigo rato solo os falta chuparselaa"	falt
"a punto de empezar la presentación del plan estratégico de grupoamas desde bankia encantados como siempre de apoyaros"	present

Tabla. Matriz de características con cinco comentarios aleatorios para FIABILIDAD

- Resultados en la matriz de características para CAPACIDAD DE RESPUESTA

Comentario	**Car. Encontrada**
" seguimos trabajando a tope para detener el desahucio de charosequeda pero si no venirse mañana a esperar a bankia a en"	trabaj
" tu experto no irá al gym por ti pero sí te ayudará en todas tus gestiones conectacontuexperto"	ayud
" y mañana pahparla también nos pide apoyo especial es el tercer intento de bankia criminal para echarla pero fátimaseq"	apoy
"por las tarjetas black y por el ere ejercido sobre las trabajadoras de bankia se merecen una dura condena txellgl b"	trabaj
"en la intranet nos han puesto calculadora para q cada empleado de bankia podamos calcular ntro impacto ambiental en lo"	emple

Tabla. Matriz de características con cinco comentarios aleatorios para CAPACIDAD DE RESPUESTA

- Resultados en la matriz de características para SEGURIDAD

Comentario	Car. Encontradas
"bankia puedo consultar el límite diario del cajero de mi tarjeta de débito desde la banca móvil"	Consult
"duro golpe del poder judicial en el caso bankia fernández ordóñez segura y retoy investigados"	segur
"debería saber q el negocio bancario se sustenta en la confianza y en los cliente"	negoci
" hoy bankia quería ejecutar charosequeda sí o sí los agentes judiciales han respondido a nuestra mediación junto a jmdp"	respond

Tabla. Matriz de características con cinco comentarios aleatorios para SEGURIDAD

- Resultados en la matriz de características para EMPATÍA

Comentario	Car. Encontradas
"hoy nos ha pasado una cosa graciosa un cliente con hipoteca en otra entidad la ha venido a reclamar a bankia pensaba q devolvíamos todas"	client
"y mañana pahparla también nos pide apoyo especial es el tercer intento de bankia criminal para echarla pero fátimaseq"	especial
" la satisfacción que da cuando llega el viernes y ves que has cumplido con toda tu agenda hora de desconectar felizviernes"	agend
"debería saber q el negocio bancario se sustenta en la confianza y en los cliente"	client
" denuncian a bankia ante inspección de trabajo por hacer trabajar fuera del horario laboral"	horari

Tabla. Matriz de características con cinco comentarios aleatorios para EMPATÍA

5.3.3. Clasificación de comentarios

Una vez que se ha obtenido las **cinco matrices de términos con las características encontradas** se procede a crear nuevas variables en cada una de ellas las cuales permiten la clasificación de comentarios. Las nuevas variables implementadas se describen a continuación:

Variable	Descripción
ID	Identificador del comentario
Porcentaje_Posibilidad (PP[16])	Promedio de acierto de que el comentario pertenezca a la dimensión en análisis PP= (NPE * 100) / NPT
Num_palabras_total (NPT)	Número total de palabras del comentario
Num_palabras_Encontradas (NPE)	Número de características encontradas

Tabla. Nuevas variables a implementar en los matrices para clasificacíon de comentarios.

Estas nuevas variables conforman el **algoritmo de clasificación** el cual nos permite agrupar los comentarios de acuerdo a la característica encontrada en los comentarios.

Ejemplo:

	PP	NPT	Carac. encontrada	NPE	infraestructura	instalaciones	oficinas	...
"Las oficinas son confortables en Bolaños de Calatrava"	12,5%	8	oficinas	1	0	0	1	0

PP = (NPE * 100)/ NPT
PP = (1 * 100) / 8
PP= 12,5 %

El resultado obtenido en el ejemplo es del 12,5% de promedio de acierto que el comentario pertenezca a la dimensión Tangibilidad.

[16] Promedio de acierto del comentario para pertenecer a una dimensión ServQual

Una consideración importante a tener en cuenta es que los comentarios en la matriz de características de cada una de las dimensiones han sido depurados, es decir se ha eliminado palabras vacías y reducido a su raíz las palabras que forman el comentario.

Por tanto el resultado real para el ejemplo anterior es:

	PP	**NPT**	Carac. encontrada	**NPE**	infraestructura	instalaciones	oficinas	…
"oficin confortable bolaños calatrava"	11,10%	4	oficinas	1	0	0	1	0

PP = (NPE * 100)/ NPT
PP = (1 * 100) / 4

PP= 25%

Otra consideración a tomar en cuenta es que, - como se observa en el apartado de diccionario de datos - existen ciertas características que están presentes en varias dimensiones como por ejemplo *"empleados"*, por lo que el **algoritmo de clasificación** calcula el promedio de aciertos PP en base a todas las características encontradas en el comentario, es decir **mientras más características encuentre en el comentario más promedio de acierto tendrá de pertenecer a un grupo.**

Por tanto para poder obtener los cinco grupos de comentarios se realiza un **filtrado de valores PP** en cada ciclo de pruebas el cual de acuerdo a su valor nos permite discriminar comentarios que no están orientados a la calidad de servicio.

Pruebas

Para las distintas pruebas se ha usado la **matriz de Fiabilidad** con 19203 observaciones para evitar el sobreajuste al usar siempre los mismos datos.

Prueba 1
Entradas: • PP>5%

Resultados

N° de observaciones agrupadas: 8746

Porcentaje de acierto mínimo: 7,7%

Promedio de acierto del grupo: 16%

Prueba 2
Entradas: • PP>20%

Resultados

N° de observaciones agrupadas: 1576

Porcentaje de acierto mínimo: 20%

Promedio de acierto del grupo: 26%

Prueba 3
Entradas: • PP>40%

Resultados

N° de observaciones agrupadas: 576

Porcentaje de acierto mínimo: 40%

Promedio de acierto del grupo: 49%

Se observa en los resultados de las tres pruebas que el número de observaciones baja cuando PP sube, esto quiere decir que el algoritmo de clasificación ha eliminado comentarios que pertenecen a otras dimensiones y que no tratan sobre la calidad de servicio.

Por tanto se ha implementado los parámetros de entrada PP>40% en las cuatro matrices restantes creando de esta manera los **cinco grupos de comentarios** que representan a las cinco dimensiones SERVQUAL

5.4.Polaridad

5.4.1. Algoritmo de búsqueda

La última fase del proyecto es obtener la polaridad asociada al conjunto de características encontradas en cada comentario de acuerdo a su agrupación mediante un algoritmo de búsqueda de sentimientos positivos y negativos

La polaridad[17] se refiere a *"la presencia o ausencia de partículas gramaticales que realizan la negación"* y dentro de una oración o comentario puede ser positiva o negativa.

Al ser los comentarios de Twitter lenguaje natural se hace imprescindible encontrar técnicas que permitan procesarlo.

Ya en anteriores fases se ha utilizado técnicas para manejar lenguaje natural como el vector de palabras el cual tenía inconvenientes con las Stop Words que debían ser tratadas, por tanto se ha utilizado una nueva técnica denominada **N-gramas** los cuales capturan información relativa al orden de las palabras de un comentario.

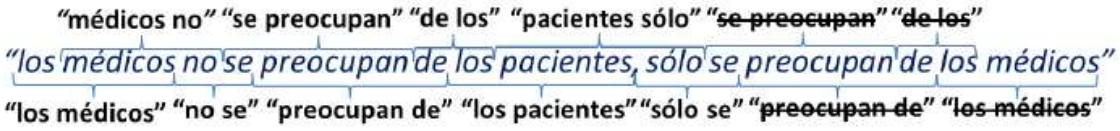


Figura 48. Ejemplo de N-gramas. Tomado de: http://eprints.ucm.es/39524/1/memoriaTFM_sergio_rincon_garcia.pdf

Por tanto el algoritmo de búsqueda de sentimientos se ha basado en la técnica de N-gramas utilizando la librería RWeka[18] de RStudio. Además se ha creado un **corpus de palabras positivas y un corpus de palabras negativas (ver anexos)** para cruzar información con el grupo de comentarios correspondiente.

Los pasos que sigue el algoritmo de búsqueda para obtener los sentimientos asociados a las características son:

1. Se particiona el comentario de acuerdo a el número de n-gramas parametrizado.

   Ejemplo:

   Comentario:

   "bbvaesp cuando mas necesitas la aplicación de mi gestor menos me funciona"

   Característica encontrada: *funciona*

   - n-gramas=1

   "aplicacion" - "bbvaesp" - "cuando" - "funciona" - "gestor" - "mas" "menos" - "necesitas"

   - *n-gramas=2*

   "aplicacion de" - "bbvaesp cuando" - "cuando mas" - "de mi" "gestor menos" - "la aplicacion" - "mas necesitas" - "me funciona" "menos me" - "mi gestor" - "necesitas la"

[17] https://es.wikipedia.org/wiki/Polaridad_(gram%C3%A1tica)
[18] https://cran.r-project.org/web/packages/RWeka/index.html

- *n-gramas=3*

"aplicacion de mi" - "bbvaesp cuando mas" - "cuando mas necesitas"
"de mi gestor" - "gestor menos me" - "la aplicacion de" - "mas necesitas la"
"menos me funciona" - "mi gestor menos" - "necesitas la aplicacion"

2. Se filtra los ngramas que contenga la característica encontrada en el comentario

   n-gramas=3
   Resultado: *"menos me funciona"*

3. Se busca en el ngrama el sentimiento asociado a la característica *"funciona"*

   Ngrama con característica encontrada: *"menos me funciona"*

   Sentimiento o palabra encontrada: *"menos"*
   Grupo al que pertenece la palabra: Negativas.

4. Se contabiliza el número de palabras positivas y el número de palabras negativas encontradas.

   (+) TOTAL POSITIVAS: 0 (-) TOTAL NEGATIVAS: 1

5.4.1.1. Pruebas y evaluaciones

Se ha realizado una serie de pruebas modificando la cantidad de n-gramas que particionará cada comentario para después evaluar la cantidad de palabras o sentimientos positivos y negativos encontrados, por lo que a mayor cantidad de sentimientos encontrados, mejor será el modelo.

N° ngrama	Tangibilidad	Fiabilidad	Cap. Respuesta	Seguridad	Empatía
1	0	0	0	0	0
2	79	155	88	182	179
3	96	167	103	205	193

Tabla. Número de palabras positivas o negativas encontradas

Evaluación

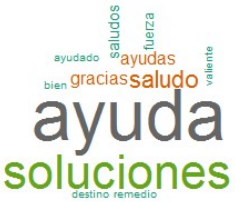
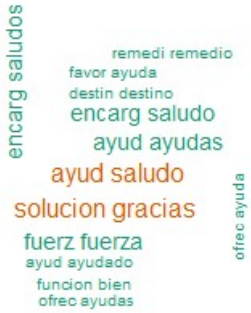
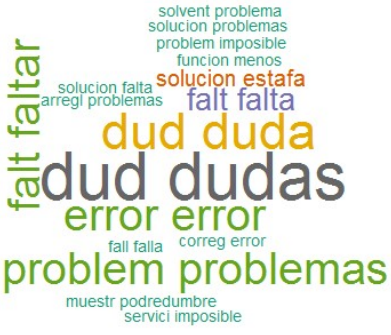
Como se observa en la tabla anterior al utilizar n-gramas de 3 el algoritmo de búsqueda encuentra más palabras positivas y negativas, por lo tanto el mejor modelo pertenece a la prueba 3.

5.5. Resultados generales

Tangibilidad

POSITIVO	
General	**Asociado a característica**
	
NEGATIVO	
General	**Asociado a característica**
	

Fiabilidad

POSITIVO	
General	**Asociado a característica**
	
NEGATIVO	
General	**Asociado a característica**
	

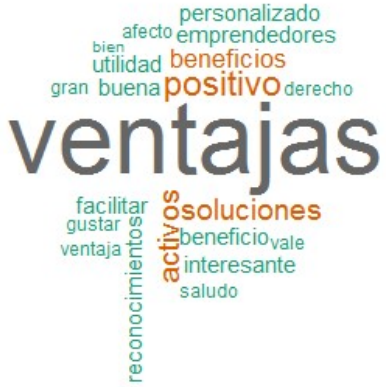
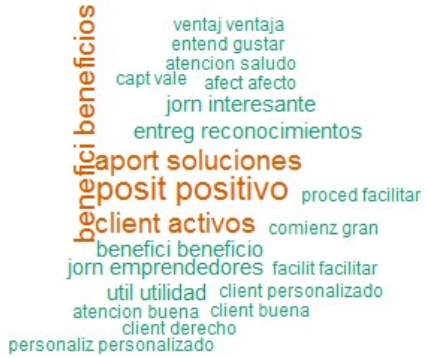
Capacidad de respuesta

POSITIVO	
General	**Asociado a característica**
	
NEGATIVO	
General	**Asociado a característica**
	

Seguridad

POSITIVO	
General	**Asociado a característica**
	
NEGATIVO	
General	**Asociado a característica**
	

Empatía

POSITIVO	
General	**Asociado a característica**
	
NEGATIVO	
General	**Asociado a característica**
	

6. Conclusiones y trabajos futuros

Las entidades financieras están compitiendo no sólo con organizaciones internas y externas en las condiciones globales de hoy por lo que, en este contexto, es importante lograr un servicio congruente, deseable y calificado ya que la calidad se logra cuando se satisfacen los deseos y expectativas de los consumidores.

En la actualidad, se sabe que medir la satisfacción del cliente es clave para desarrollar estrategias orientadas al consumidor. Sin embargo, existen menos acuerdos sobre el desarrollo de metodologías y escalas uniformes para medir la calidad del servicio. Si bien es cierto que la escala SERVQUAL ha tenido más éxito que otras iniciativas en este campo, las diversas adaptaciones y cambios introducidos en las escalas de medición a menudo dificultan la comparación de los resultados con el tiempo lo cual es un aspecto clave que las empresas deben tener en cuenta al implementar sus estrategias orientadas al mercado.

En cuanto al proyecto, se ha presentado un trabajo cuya finalidad es ofrecer una solución a un problema real desde un punto de vista de minería de datos. Este problema es un caso particular de clasificación y análisis de sentimientos por lo que está dentro de la categoría de Minería de Textos. Dicho problema es aún un problema abierto dentro del área de investigación del Procesamiento del Lenguaje Natural.

En particular, se trata de medir la calidad del servicio en la Banca Española con una escala como SERVQUAL de manera automática, usando como fuente de información las opiniones de la red social Twitter. Esto permitirá ahorrar recursos destinados a recopilar información y generar encuestas, además de estandarizar dicha medición.

Por lo que, la solución que ofrece el presente proyecto propone lo siguiente:

Clasificación de comentarios basados en características encontradas que permitan definir a cuál de las cinco dimensiones SERVQUAL pertenece la opinión.

Búsqueda de sentimientos asociados a cada característica encontrada para determinar si la opinión es positiva o negativa.

En cuanto al problema de la clasificación de comentarios ha requerido sortear el obstáculo de tener muchas opiniones "basura" que no necesariamente reflejaban una opinión sobre el servicio ofrecido por un determinado Banco. Para ello se propuso crear un diccionario de características. Sus resultados han sido favorables, pero se requiere refinar más dichos diccionarios para tener agrupaciones de comentarios más fiables.

También se ha propuesto la creación de matrices de términos las cuales ha permitido cruzar la información entre los comentarios y los diccionarios de características para obtener las cinco agrupaciones de comentarios que representan las cinco dimensiones SERVQUAL. Los resultados obtenidos han sido muy favorables ya que por este medio, se pudo encontrar de una manera rápida las características SERVQUAL en cada uno de los comentarios. Sin embargo se requirió de una depuración inicial usando técnica de Minería de Texto para que le modelo de clasificación obtenga buenos resultados.

En cuanto al problema de la búsqueda de sentimientos asociados ha requerido sortear el obstáculo de no tener librerías en R que permitan obtener la polaridad de un comentario escrito en lenguaje español. Para ello se propuso la creación de un corpus de palabras positivas y otro corpus de palabras negativas. Sus resultados no han sido del todo favorables ya que existían palabras comunes tanto en estos corpus como en los diccionarios de características que hacían confuso la obtención del sentimiento asociado, por lo que se requiere una mejor depuración en los corpus de palabras positivas y negativas.

También se propuso la utilización de la técnica de n-gramas para obtener el sentimiento asociado a la característica arrojando resultados muy favorables. Esto permitió realizar comparación de resultados entre las distintas formas de n-gramas resultando como ganador el uso de los trigramas.

En cuanto a los resultados obtenidos tanto de clasificación como sentimientos asociados se puede decir que:

En la dimensión de Tangibilidad la percepción positiva del cliente es que el equipamiento es excelente sin embargo la percepción negativa es el difícil manejo de los servicios bancarios. De manera general se podría decir que los bancos deben mejorar el manejo de los distintos puntos de contacto que ofrecen, ejemplo: Portales Web.

En la dimensión de Fiabilidad, la percepción positiva del cliente es que se presta ayuda y da soluciones, sin embargo la percepción negativa es que genera muchas dudas y lo consideran

una estafa. En balance general en esta dimensión la Banca debe mejorar considerablemente ya que los clientes lo consideran poco fiable.

En la dimensión de Capacidad de Respuesta la percepción positiva es que se brinda apoyo oportuno y los empleados son solidarios. Para la percepción negativa es evidente que existen dudas respecto al servicio brindado. Por tanto en esta dimensión la Banca debe reforzar su apoyo oportuno para eliminar dudas del servicio.

En la dimensión de Seguridad la percepción positiva es que los clientes sienten seguridad y compromiso del trabajador. Para la percepción negativa también se refleja que existen dudas respecto a los conocimientos del empleado. Esta es una de las dimensiones con la mejor percepción positiva por parte del cliente.

En la dimensión de Empatía es claro que la percepción positiva se basa en las ventajas que ofrece los servicios además de los beneficios, el emprendimiento, una buena atención y clientes satisfechos. No se tiene perspectivas negativas para esta dimensión. Esta es sin lugar a duda el punto más fuerte que tiene la Banca Española en cuanto a la percepción del servicio por parte de sus clientes.

Por último, en cuanto a los resultados del proyecto, han sido razonablemente buenos, dado que no había referencias de resultados anteriores más allá de lo que se explicó en el estado del arte, debido a que es un problema real sin resolver.

Ahora bien en el buscador de sentimientos asociados quedan muchas ideas por explorar para mejorar sus resultados como trabajo futuro. Por ello se propone lo siguiente:

- Probar técnicas más modernas de búsqueda de características como Word2vec o uso de modelos basados en Deep Learning.
- Probar enfoques de análisis semánticos y sintácticos además de un enfoque léxico.
- Investigar más a profundidad librerías de análisis de sentimientos para frases en español con R Studio.
- Probar técnicas avanzadas de Minería de Textos para encontrar adjetivos cuantitativos que cambian el sentido de la frase. Por ejemplo, *"la no ayuda de su empleado ocasionó un disgusto en el cliente"*

7. Referencias

- Agirre, E. (2015). *Semantic Textual Similarity English, Spanish and Pilot on Interpretability*. Obtenido de http://www.aclweb.org/anthology/S15-2045
- Ganesan, K. (23 de 11 de 2014). *text-analytics101*. Obtenido de http://text-analytics101.rxnlp.com/2014/11/what-are-n-grams.html
- Legutier. (1 de Septiembre de 2005). *Legutier*. Obtenido de http://legutier.blogspot.com.es/2005/09/qu-es-categorizar-textos.html
- Molina, L. C. (s.f.). *UOC*. Obtenido de http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html
- Parasuraman, Z. (1985). A conceptual model of service quality and itsimplications for future research. *Journal of Marketing*, 41-50.

- R., L. (2009). A review or twenty years of SERVQUAL research. *International Journal of Quality and Service Sciences*, 172-198.
- Ramón A. Carrasco, F. M.-L.-F. (2012). A model for the integration of e-financial services questionnaires with SERVQUAL scales under fuzzy linguistic modeling. *elsevier*.
- *UGR*. (s.f.). Obtenido de http://www.ugr.es/~bioestad/_private/cpfund10.pdf
- Sergio García (2015). Minería de Textos y Análisis de Sentimientos en sanidadysalud.com

Web

- Deep Learning: https://www.unocero.com/2013/08/22/aprendizaje-profundo-aplicado-al-lenguaje-natural/
- SERVQUAL Calidad de servicio: https://www.aiteco.com/modelo-servqual-de-calidad-de-servicio/
- Similaridad: https://es.oxforddictionaries.com/definicion/similaridad
- SERVQUAL Preguntas de cuestionario: https://rodas5.us.es
- Twitter Definición de hashtag: https://www.significados.com/hashtag/
- MONGO DB. Fundamentos: https://www.mongodb.com/es
- Minería de Texto. Matriz de términos: https://mukom.mondragon.edu/ict/mineria-de-textos-para-la-clasificacion-de-documentos-en-espanol-con-r/
- Minería de Texto. Matriz de términos: https://nlp.stanford.edu
- Minería de Texto. Stop Words: http://www.vozidea.com/que-son-las-stop-words-o-palabras-vacias
- Minería de Texto. Redución de términos a su raíz: http://www.cemla.org/PDF/boletin/PUB_BOL_LXII-03-01.pdf
- Minería de Texto Polaridad :https://es.wikipedia.org/wiki/Polaridad_(gram%C3%A1tica)

Paquetes R

- TwitterR: https://cran.r-project.org/web/packages/twitteR/twitteR.pdf
- Tm: https://cran.r-project.org/web/packages/tm/tm.pdf
- Ggplot: https://cran.r-project.org/package=ggplot2
- WordCloud: https://cran.r-project.org/package=wordcloud
- SnowballC:https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf
- Rweka: https://cran.r-project.org/package=RWeka

8. Anexos

Activación de cuentas en Twitter

Nuestra fuente de información principal fue proporcionada por la red social Twitter la cual para acceder a ella se debe seguir una serie de pasos que se enumeran a continuación:

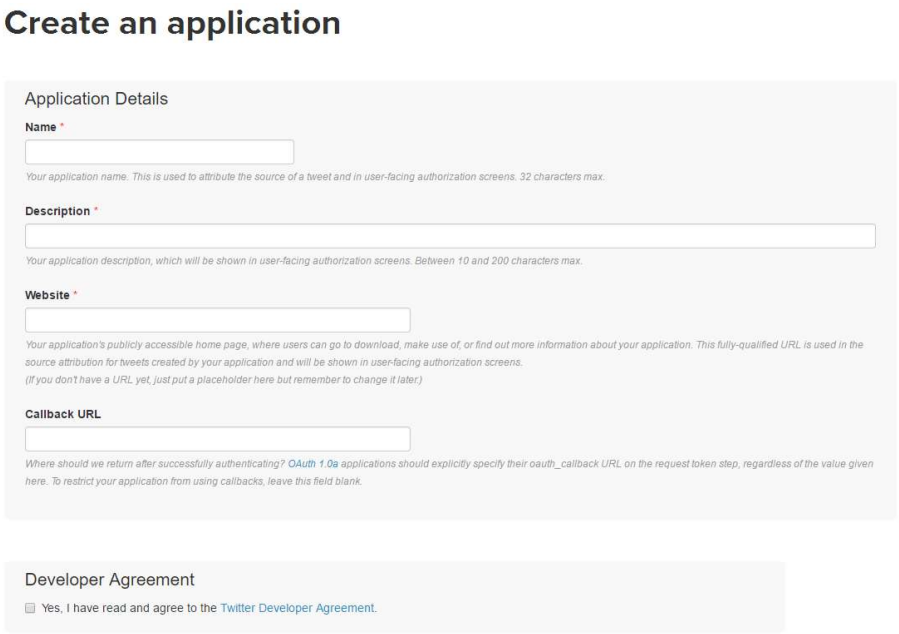
- Activar una cuenta de Twitter como desarrollador[19]

Para activar una cuenta como desarrollador se debe:

  - Tener una cuenta de Twitter.
  - Registrar un número telefónico en la cuenta de usuario de Twitter.

- Crear una API de Twitter

Una vez que se ha activado la cuenta como desarrollador se puede crear aplicaciones que utilicen la API[20] de twitter. Para ello se debe llenar un formulario para crear una aplicación.



Como se puede observar en la figura los campos a llenar en el formulario son:

Name: Nombre de la APP. En nuestro caso el nombre escogido es: "t_sentiment"

Description: Descripción de la app de twitter. Ejemplo: "Api para búsquedas enTwitter"

[19] https://apps.twitter.com/
[20] Abreviatura de "Interfaz de Programación de Aplicaciones"

WebSite: Página Web de inicio del aplicativo que va a usar el Api de Twitter, en nuestro caso al ser una aplicación de escritorio[21] no hace falta su registro.

Callback URL: Página web de retorno una vez que se haya iniciado sesión en el API de Twitter, en este caso no es necesario su registro

Developer Agreement: Acuerdo de desarrollador. Registro obligatorio de dicho campo que notifica las limitaciones, términos y condiciones de uso de la API de Twitter.

Una vez registrada la información necesaria en el formulario se crea el API de Twitter como se muestra en la siguiente figura:

Figura. API de Twitter creada

- Obtener las credenciales de acceso a la API de Twitter

Una vez creada la API de Twitter se puede obtener las respectivas credenciales de acceso como se observa en la figura:

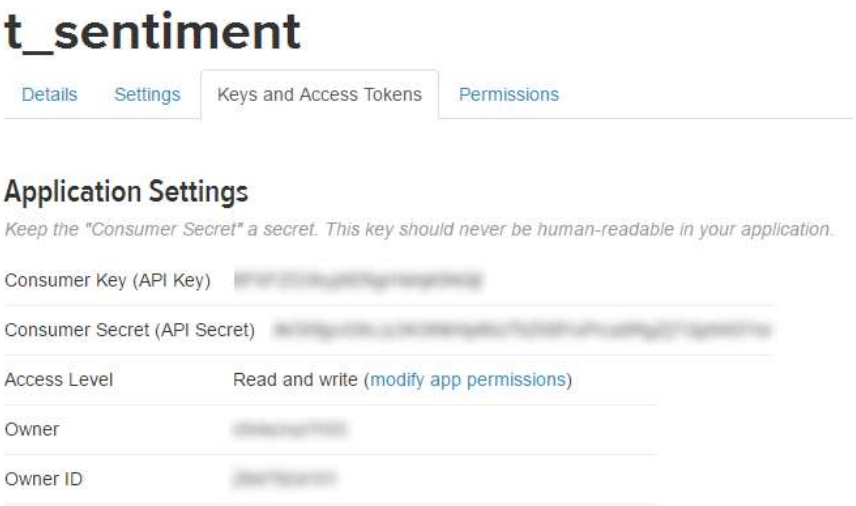


Figura. Credenciales para acceder al API de Twitter.

[21] "Programa que se instala o ejecuta en un ordenador" http://www.arumeinformatica.es/dudas/pagina-web-aplicacion-web-y-aplicacion-de-escritorio-cual-es-la-diferencia/

Cuestionario servQual

Dimensión	Preguntas
TANGIBILIDAD	1. X cuenta con un equipamiento de aspecto moderno.
	2. Las instalaciones físicas de X son visualmente atractivas.
	3. Los empleados de X tienen buena presencia
	4. En X, el material asociado con el servicio (como los folletos o los comunicados) es visualmente atractivo.
FIABILIDAD	5. Cuando X promete hacer algo en una fecha determinada, lo cumple.
	6. Cuando tiene un problema, X muestra un interés sincero por solucionarlo.
	7. X lleva a cabo el servicio bien a la primera.
	8. X lleva a cabo sus servicios en el momento que promete que va a hacerlo.
	9. X pone énfasis en unos registros exentos de errores.
CAPACIDAD DE RESPUESTA	10. Los empleados de X le comunican con exactitud cuándo se llevarán a cabo los servicios.
	11. Los empleados de X le proporcionan un servicio rápido.
	12. Los empleados de X siempre están dispuestos a ayudarle.
	13. Los empleados de X nunca están demasiado ocupados para responder a sus preguntas.
SEGURIDAD	14. El comportamiento de los empleados de X le inspira confianza.
	15. Se siente seguro en sus transacciones con X.
	16. Los empleados de X suelen ser corteses con usted.
	17. Los empleados de X tienen los conocimientos necesarios para contestar a sus preguntas.
EMPATÍA	18. X le proporciona atención individualizada.
	19. X tiene unos horarios de apertura o atención adecuados para todos sus clientes.
	20. X cuenta con unos empleados que le proporcionan una atención personalizada.
	21. X se interesa por actuar del modo más conveniente para usted.
	22. Los empleados de X comprenden sus necesidades específicas.

Stop Words en español

a	alguna	aquella	añadió	cierto	cosas	cuántas	demas	dieron	embargo
actualmente	algunas	aquellas	aún	ciertos	creo	cuánto	demasiada	diferente	empleais
acuerdo	alguno	aquello	b	cinco	cual	cuántos	demasiadas	diferentes	emplean
acá	algunos	aquellos	bajo	claro	cuales	cómo	demasiado	dijeron	emplear
adelante	algún	aqui	bastante	comentó	cualquier	d	demasiados	dijo	empleas
ademas	alli	aquél	bien	como	cualquiera	da	demás	dio	empleo
además	allá	aquélla	breve	con	cualquieras	dado	dentro	donde	en
adrede	allí	aquéllas	buen	conmigo	cuan	dan	deprisa	dos	encima
afirmó	alrededor	aquéllos	buena	conocer	cuando	dar	desde	durante	encuentra
agregó	ambos	aquí	buenas	conseguimos	cuanta	de	despacio	día	enfrente
ahi	ampleamos	arriba	bueno	conseguir	cuantas	debajo	despues	días	enseguida
ahora	antano	arribaabajo	buenos	considera	cuanto	debe	después	dónde	entonces
ahí	antaño	aseguró	c	consideró	cuantos	deben	detras	e	entre
ajena	ante	asi	cabe	consigo	cuatro	debido	detrás	ejemplo	era
ajenas	anterior	así	cada	consigue	cuál	decir	dia	el	eramos
ajeno	antes	atras	casi	consiguen	cuáles	dejar	dias	ella	eran
ajenos	apenas	aun	cerca	consigues	cuán	dejó	dice	ellas	eras
al	aproximadamente	aunque	cierta	contigo	cuándo	del	dicen	ello	eres
algo	aquel	ayer	ciertas	contra	cuánta	delante	dicho	ellos	es

esa	estoy	g	hacerlo	incluso	lado	me	momento	ni	nunca
esas	estuvo	general	haces	indicó	largo	mediante	mucha	ningun	o
ese	está	gran	hacia	informo	las	medio	muchas	ninguna	ocho
eso	están	grandes	haciendo	informó	le	mejor	mucho	ningunas	os
esos	etc	gueno	hago	intenta	lejos	mencionó	muchos	ninguno	otra
esta	ex	h	han	intentais	les	menos	muchísima	ningunos	otras
estaba	excepto	ha	hasta	intentamos	llegó	menudo	muchísimas	ningún	otro
estaban	existe	haber	hay	intentan	lleva	mi	muchísimo	no	otros
estado	existen	habia	haya	intentar	llevar	mia	muchísimos	nos	p
estados	explicó	habla	he	intentas	lo	mias	muy	nosotras	pais
estais	expresó	hablan	hecho	intento	los	mientras	más	nosotros	para
estamos	f	habrá	hemos	ir	luego	mio	mí	nuestra	parece
estan	fin	había	hicieron	j	lugar	mios	mía	nuestras	parecer
estar	final	habían	hizo	jamás	m	mis	mías	nuestro	parte
estará	fue	hace	horas	junto	mal	misma	mío	nuestros	partir
estas	fuera	haceis	hoy	juntos	manera	mismas	míos	nueva	pasada
este	fueron	hacemos	hubo	k	manifestó	mismo	n	nuevas	pasado
esto	fui	hacen	i	l	mas	mismos	nada	nuevo	paìs
estos	fuimos	hacer	igual	la	mayor	modo	nadie	nuevos	peor

pero	por	pueden	quiénes	sean	sin	supuesto	tanto	toda	tu
pesar	porque	puedo	qué	segun	sino	sus	tantos	todas	tus
poca	posible	pues	r	segunda	so	suya	tarde	todavia	tuvo
pocas	primer	q	raras	segundo	sobre	suyas	te	todavía	tuya
poco	primera	qeu	realizado	según	sois	suyo	temprano	todo	tuyas
pocos	primero	que	realizar	seis	sola	suyos	tendrá	todos	tuyo
podeis	primeros	quedó	realizó	ser	solamente	sé	tendrán	tomar	tuyos
podemos	principalmente	queremos	repente	sera	solas	sí	teneis	total	tú
poder	pronto	querer	respecto	será	solo	sín	tenemos	trabaja	u
podria	propia	quien	s	serán	solos	sólo	tener	trabajais	ultimo
podriais	propias	quienes	sabe	sería	somos	t	tenga	trabajamos	un
podriamos	propio	quienesquiera	sabeis	señaló	son	tal	tengo	trabajan	una
podrian	propios	quienquiera	sabemos	si	soy	tales	tenido	trabajar	unas
podrias	proximo	quiere	saben	sido	soyos	tambien	tenía	trabajas	uno
podrá	próximo	quiza	saber	siempre	sr	también	tercera	trabajo	unos
podrán	próximos	quizas	sabes	siendo	sra	tampoco	ti	tras	usa
podría	pudo	quizá	salvo	siete	sres	tan	tiempo	trata	usais
podrían	pueda	quizás	se	sigue	sta	tanta	tiene	través	usamos
poner	puede	quién	sea	siguiente	su	tantas	tienen	tres	usan

usar	vez	ésta
usas	vosotras	éstas
uso	vosotros	éste
usted	voy	éstos
ustedes	vuestra	última
v	vuestras	últimas
va	vuestro	último
vais	vuestros	últimos
valor	w	
vamos	x	
van	y	
varias	ya	
varios	yo	
vaya	z	
veces	él	
ver	ésa	